Public Finance
A Normative Theory

Third Edition

Richard W. Tresch
Department of Economics
Boston College
Chestnut Hill, Massachusetts

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD • PARIS
SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an Imprint of Elsevier

Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA
225 Wyman Street, Waltham, MA 02451, USA
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

First published 1981
Second edition 2002
Third edition 2015

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-12-415834-4

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

Typeset by TNQ Books and Journals
www.tnq.co.in

Printed and bound in the United States

Preface

The third edition of the textbook might be described as a substantial partial revision of the second edition. I added four new chapters to make the book more useful to a wider audience: two chapters on social insurance, one covering retirement pensions with an emphasis on the U.S. Social Security System and one on medical insurance; a chapter on behavioral public finance; and a chapter on international public finance, with an emphasis on international tax issues. The book could not increase in length, however, so these new chapters had to replace some of the existing chapters. I decided that the least costly chapters to delete were those on cost–benefit analysis, on the grounds that there are many good cost–benefit texts on the market and also that the cost–benefit material was tangential to my goal of presenting a comprehensive treatment of the mainstream normative theory of the public sector.

There were a few other major changes as well. I deleted the appendix to Chapter 8 on U.S. antipollution policy and reworked Chapter 8 to feature global warming as the example of a consumption–production externality. In addition to its current interest, global warming gave me an opportunity to discuss the issues associated with external effects that occur far in the future. I also added two new appendices: one on the distinction between the external and internal margins in response to tax and transfer policies, using Emmanuel Saez's seminal article as the focal point, and the other on tax reform. The latter considers what tax theory has to say about four broad-based tax reform proposals that are common in the economics literature, bringing together some theoretical results from previous chapters and then turning to other issues that are not discussed elsewhere in the text. These include whether income from capital should be taxed, Michael Kremer's call for age-based marginal tax rates, and the general problem of commitment in second-best analysis with imperfect information. The new appendices are in keeping with an attempt to bring more empirical analysis into the text wherever I could. I had to be quite selective here, to adhere to the normative thrust of the text and to keep the length of the text within bounds. Some examples include evidence on the deficits of the urban rail transit systems in the United States to buttress the conjecture that they represent hard-case decreasing services (Chapter 9), newer evidence on the Martin Feldstein's proposed efficiency measure of the elasticity of taxable income (Chapter 13), and evidence on the degree of Tiebout sorting in the United States (Chapter 27).

A final change of note is that the text will now come with an accompanying Web site that will include PowerPoint slides for each chapter and end-of-chapter questions. I welcome suggestions from readers for other useful material to include (perhaps brief accounts of new topics as they appear in the literature?).

The changes in the text notwithstanding, users of the second edition should feel quite at home with the third edition. The core chapters are largely unchanged, other than adding some empirical material. These include the introductory chapters; the public expenditure chapters covering externalities, decreasing cost services, and transfer payments; the tax chapters on efficiency, equity, and tax incidence; the analysis of taxes and transfers under asymmetric information; and the three chapters on federalism. As before, I begin with the first-best analysis of both public expenditure and tax theory, followed by the second-best analysis of each theory. The emphasis throughout continues to be on the mainstream normative theory of the public sector, and is almost all micro-oriented. Finally, the level of mathematical analysis remains the same as before, suitable for both PhD and Masters programs, and even mathematically oriented undergraduate programs.

In closing, I want to take the opportunity that the Preface affords to thank a number of people. First and foremost is Peter Diamond, who taught the public sector course when I was a graduate student at MIT and motivated me to specialize in the public sector. I have acknowledged in previous editions his influence on my presentation of the core public expenditure and tax theory in the text. His influence continues in some of the newer materials such as social insurance and tax reform. I want to acknowledge again my gratitude to Nan Friedlaender for her support and mentoring when I began my career at Boston College. My academic formation also owes much to William Rhoads, who taught the public sector course when I was at Williams College, and Anthony Davidowski, who taught me mathematics for 3 years in high school. I learned quite a bit about how to teach from both of them, and the obvious joy they derived from teaching no doubt influenced me to become an academic. A final heartfelt thank you goes out to

all the fine people at Academic Press for their expertise, help, and encouragement in producing the third edition, particularly my editor J. Scott Bentley, editorial project managers Melissa Murray and Mckenna Bailey, copyediting project manager Lisa Jones, and marketing manager Cindy Minor.

Richard W. Tresch
Chestnut Hill, MA, USA

Chapter 1

Introduction to Normative Public Sector Theory

Chapter Outline

Public sector economics is the study of government economic policy. Its primary goal is to determine whether government policies promote a society's economic objectives. This happens to be quite an ambitious goal. The advanced Western market economies experienced enormous growth in the size and influence of their government sectors during the last half of the twentieth century, and economic analysis of the public sector has reflected this growth. No single textbook on public sector economics can possibly hope to capture the variety and richness of the professional economic literature on government policy, even at an introductory level. Consequently, a public sector text must begin by defining its limits.

We have chosen to limit both the subject matter and the approach of this text. The text concentrates on the microeconomic theory of the public sector in the context of capitalist market economies. The macroeconomic theory of the public sector, commonly referred to as fiscal policy, receives little attention. In addition, the text focuses on the normative theory of the public sector rather than the positive theory. The normative theory considers what governments ought to be doing in accordance with norms that are broadly accepted by a society. In contrast, the positive theory of the public sector emphasizes the incentives generated by existing governmental institutions and policies and their resulting economic effects, without necessarily judging their effectiveness in terms of some accepted norms. A complete separation of normative and positive theory is impossible, of course. A normative analysis must make assumptions about how agents will respond to various government polices; otherwise, it cannot predict whether a given policy will achieve particular norms. Therefore, the text pays some attention to the empirical literature on the responses to government policies, for example, how the supply of labor responds to income taxation. In every chapter, though, our primary emphasis is on the normative theory of government policy under standard assumptions about economic behavior, such as utility maximization by consumers and profit maximization by producers.

That a consensus, mainstream, normative theory of the public sector should have evolved at all in Western economic thought is perhaps surprising, yet there is remarkable agreement on the problems the government ought to address and the appropriate course of government action in solving them. The consensus has arisen in part because the vast majority of Western public sector economists embrace the same set of policy norms, even though their political tastes may vary along the entire liberal–conservative

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00001-7

spectrum. In addition, most public sector economists have chosen the same basic model to analyze all public sector economic problems. Given the same norms and a common analytical framework, consensus was inevitable.

The only serious competitors to the mainstream view of the public sector are the theory of public choice and behavioral economics. James Buchanan was the founding father of the theory of public choice, for which he received the Nobel Memorial Prize in Economics. He garnered an enthusiastic following, and his public choice perspective has been influential in policy analysis. Public choice remains a distinctly minority view, however, and its approach is more positive than normative. For these reasons, this text considers the public choice perspective only when it has been especially influential in challenging mainstream positions.

Behavioral economics is a newer competitor to the mainstream theory. It attempts to apply psychological principles to help understand behavior that is otherwise at odds with the mainstream assumption that people act to maximize their own self-interests. It is gaining momentum within the profession in all areas of economics, enough so that we have devoted a chapter to explore some of its positive and normative implications for public sector theory. It is far from ready to supplant the mainstream economic theory of the public sector, however.

The first three chapters introduce the mainstream normative theory of the public sector. Chapter 1 begins by describing the four fundamental questions that a normative analysis must address and shows how a particular set of values or norms shared by virtually all Western economists has produced a consensus on how to answer them. The chapter also introduces the public choice perspective on the appropriate economic role of the government.

Chapter 2 presents a baseline "textbook" version of the basic general equilibrium model that is used to develop normative public sector decision rules. The chapter emphasizes how the norms described in Chapter 1 are incorporated into the formal model.

Chapter 3 concludes the introductory material with two methodological points. The first point is the distinction between first-best and second-best analyses. First-best analysis assumes that a government is free to pursue whatever policies are necessary to reach society's economic goals. It is restricted only by the two natural fundamentals inherent in any economy: individuals' preferences over goods and factor supplies and the available production technologies for turning inputs into outputs. Second-best theory assumes, more realistically, that a government is constrained beyond the two fundamentals in pursuing society's goals. For example, a government may lack the information it needs about individuals' preferences or production technologies to design first-best policies, or it may be forced to use certain kinds of taxes that distort economic decisions.

The second methodological point relates to the political content of the baseline general equilibrium model developed in Chapter 2. The discussion centers on the general impossibility theorem of Kenneth Arrow, another Nobel Laureate in economics. Arrow's theorem, which he published in 1951, stands as one of the landmark results of twentieth-century political philosophy (Arrow, 1951). He proved that, in general, the political decisions needed to achieve any social objective, economic or otherwise, cannot be made in a manner that would be acceptable to a democratic society. This was a devastating blow to the concept of a democratic or representative government. Any normative economic theory of the public sector must acknowledge the huge political shadow cast over it by Arrow's theorem.

THE FUNDAMENTAL NORMATIVE QUESTIONS

A normative economic theory of the public sector addresses four fundamental questions:

1. The primary normative question, upon which all others turn, is the question of *legitimacy*: In what areas of economic activity can the government legitimately become involved? The legitimacy question points to the expenditure side of government budgets, asking what items we should expect to find there and why.
2. Once the appropriate sphere of government activity has been determined, the next question concerns how the government should proceed. What *decision rules* should the government follow in each area?

Taken together, these two questions comprise the heart of normative public sector theory, commonly referred to as the theory of government expenditures.

3. The theory of government expenditures in turn suggests a third normative question: How should the government finance these expenditures? Analysis of this question provides the basis for a comprehensive *normative theory of taxation* (more generally, a theory of government revenues). The theory of taxation is not necessarily distinct from the theory of government expenditures, however. Frequently, the decision rules for government expenditures incorporate taxes as part of the solution. When this occurs, the theory of taxation is effectively subsumed within the theory of government expenditures. A common example is the use of taxes to correct for externalities. Often, however, expenditure theory does not specify a payment mechanism for financing particular expenditures, in which case the theory of taxation takes on a life of its own. For example, broad-based taxes such as the federal and state personal income taxes are used to finance a number of different

expenditures. The design of these taxes depends on norms developed specifically to address the problem of how general tax revenues should be collected.

4. The fourth normative question arises in the context of a federalist system of governments. A federalist system is a hierarchical structure of governments in which each citizen is, simultaneously, a member of more than one governmental jurisdiction. The United States, with its national government, 50 state governments, and over 89,000 local government entities is but one example. Most countries have a federalist structure.

Having determined the legitimate areas of government activity in answering the first question, the *theory of fiscal federalism* raises two additional questions, both in the nature of assignment or sorting problems. The first concerns the assignment of functions throughout the fiscal hierarchy: Which tasks should each government perform? The second concerns the sorting of people within the fiscal hierarchy: Where should each person live?

A society must assign the legitimate functions of government among the various levels of government so that public policies do not work at cross-purposes in pursuing economic objectives. One can easily imagine potential conflicts arising without proper coordination, such as one government heavily taxing one group of people while another government is simultaneously trying to transfer income to the same group, or one town actively promoting industrial development that damages the environment of neighboring towns. The theory of fiscal federalism, then, accepts as given the normative rules for public expenditures and taxation established in response to the first three questions. It merely tries to ensure that these rules are followed consistently throughout the entire fiscal structure.

The sorting of people by jurisdiction is closely related to the assignment of functions, since people choose where to live partly in response to the expenditure and tax mix in different localities. Once people choose where to live, they then become voters who influence the expenditure and tax mix within that locality. Therefore, the movement of people across localities can affect how well lower level governments perform their assigned functions or, indeed, whether they can perform certain functions at all. The assignment of functions and people are the two main issues in the normative theory of fiscal federalism.

Parts II and III of the text develop the normative theories of public expenditures and taxation under the assumption of a single government. Part IV considers the special problems associated with a federalist system of government. It also includes a chapter on the international taxation of capital, which is related to federalism in the sense that capital flows almost as easily across borders worldwide as it does across states and provinces within any one country.

GOVERNMENT EXPENDITURE THEORY: PHILOSOPHICAL UNDERPINNINGS

The answer a society gives to the first normative question on the legitimate functions of government is culturally determined. It turns on essentially the same set of cultural norms and attitudes that lead to the choice of a particular economic system.

Economic systems are typically characterized as lying along a spectrum whose end points are centrally planned socialism and market capitalism in their purest forms. All actual economic systems are mixtures of the two. The four principal characteristics of pure centrally planned socialism are centralized economic decision making undertaken by a bureau of the national government, the use of a national plan developed by the central bureau to process all relevant economic information and coordinate economic exchanges, public ownership of capital and possibly land as well, and the use of moral suasion to motivate agents to carry out the national plan "for the good of the state." The four principal characteristics of pure market capitalism are decentralized economic decision making undertaken by individuals and firms, the use of markets to process all the relevant information that agents need to engage in exchange and to coordinate their economic exchanges, private ownership of capital and all other resources, and the use of material rewards to motive agents to engage in exchange. A society's view of the legitimate functions of government clearly depends upon whether it has chosen an economic system closer to centrally planned socialism or to market capitalism.

Humanism, Consumer Sovereignty, Capitalism, and the Government

The normative economic theory of the public sector that developed in the West is closely tied to market capitalism. This is hardly surprising, as all the developed market economies in the West are positioned much closer to the capitalist end of the economic spectrum than to the socialism end of the spectrum. On a more basic level, however, the seeds of the preference for capitalism itself were planted when humanism swept through Europe in the fifteenth century and spawned the Reformation. Humanism was the philosophical revolution that replaced the quest for the divine with the quest for individual development and well-being as the central purpose of human endeavor. Among other things, humanism established the principle of *consumer sovereignty* (and producer sovereignty) as a fundamental value judgment or norm in the conduct of economic affairs. The principle states that consumers (producers) are the best judges of their own well-being and should be allowed to pursue their self-interests toward this end. The decentralized nature of market capitalism, coupled

with the private ownership of property, gave individuals (and firms) the freedom to pursue their self-interests. From a humanistic perspective, then, decentralization and private property are powerful attractions of capitalism, whatever other economic properties capitalism might possess. Likewise, the mainstream public sector theory became closely tied to market capitalism in the West because it, too, is rooted in humanism and takes the principle of consumer (and producer) sovereignty as a fundamental value judgment. The same can be said of any branch of Western economic theory—consumer economics, industrial organization, international trade, and so forth. Mainstream Western economists are all children of humanism.

The humanistic foundation of public sector theory has produced a consensus among Western economists on three issues related to the role of government in the economy: the legitimate functions of government, the appropriate goals of public policy, and how the government should proceed in pursuing the goals. In other words, there is broad agreement on the answers to the first two fundamental questions of the normative theory, the questions that comprise the theory of public expenditures.

The Legitimate Functions of Government

The government's economic role, broadly speaking, is to enhance the performance of the market economy. The market always takes precedence for solving agents' economic problems and allocating resources, and a perfectly competitive market economy is accepted as the ideal economic system. But even a perfectly competitive economy cannot solve all economic problems, and many markets are far from perfectly competitive. The government, therefore, has a legitimate role to play in a market economy.

Government activity gains its legitimacy through market failure. The government should perform those economic functions that markets cannot perform at all or that markets perform badly enough to warrant government intervention. Reasonable people may disagree in particular instances on whether the market is performing "badly enough" to justify government intervention, but market failure is always the test. Government activity is never justified if markets are performing adequately. Despite the room for disagreement, there happens to be fairly broad agreement on the list of legitimate government functions implied by the market failure criterion. We will consider them below.

The Goals of Government Policy

The goal of any economic system is often loosely stated as promoting the economic well-being of a nation's citizens, in keeping with the humanist philosophy. The same goal applies to government policy as well. This goal is difficult to define more precisely, however. It cannot be to maximize each individual's economic well-being or even to allow individuals to reach their full economic potential. These goals may sound attractive, but they are meaningless because they violate the Law of Scarcity; only a limited amount of resources are available to promote each individual's economic well-being or economic potential. Therefore, Western economists have chosen two proximate goals that are directly related to individual well-being as the principal economic objectives: efficiency and equity (fairness). When economists speak of promoting the "public interest," they mean the public's interest in efficiency and equity.

Efficiency

The efficiency criterion is the standard one of *pareto optimality* stated in terms of people: An allocation is efficient if it is impossible to reallocate resources such that one person can be made better off without making at least one other person worse off. Moreover, the people themselves must be the judges of whether they are better or worse off, by the principle of consumer sovereignty. An immediate corollary is that the government should pursue all *pareto-superior* allocations, those that make at least one person better off without making anyone else worse off.

Equity

The equity criterion is more difficult to define because neither economists nor anyone else has reached a consensus on what is equitable or fair in the realm of economic affairs. About all one can point to are some notions of equity that commonly appear in the economic literature. They fall into two categories: process equity and end-results equity. *Process equity* is a judgment about the rules of the economic game: Are the rules fair, independently of the outcomes that result? *End-results equity* is a judgment about the outcomes of the economic game: Are the outcomes fair, independently of how they were achieved?

Process Equity

One widely held norm of process equity is *equal opportunity*, or equal access, which says that all people should be allowed to pursue whatever opportunities they are willing and able to pursue. Equal opportunity rules out inappropriate forms of discrimination, such as denying people access to certain jobs on the basis of their race, religion, or sex. Another widely held norm of process equity is *social mobility*, which refers to the ability of individuals or families to move within the distribution of income or wealth over time. The antithesis of social mobility is the caste system, in which people are born into a certain position within the distribution and must remain there for life.

One of the great attractions of a market economy is that it fosters both equal opportunity and social mobility so long as markets are competitive.

The call for process equity is most closely associated with the philosopher Robert Nozick, who believes that equity begins and ends with the rules of the game.[1] He argues that any outcome of a fair game is fair. In particular, if the rules of the economic "game" are fair, then any outcome the economy generates is inherently fair. Societies have tended to reject Nozick's view on economic matters, however. Nations routinely make independent judgments about outcomes, especially about the extremes of poverty and wealth. They have been willing to transfer resources to the poor in cash and in kind to ease the burden of poverty, paid for by taxes on the nonpoor. President Lyndon Johnson went so far as to declare a war on poverty in 1964 with the intent of eradicating poverty within the United States, a war that is far from being won.

The majority of economists worry about end-results equity as well. One reason why may be that the rules governing the game are commonly seen to be inherently unfair. Think of the game as a race to economic well-being run within the confines of a market economy. The problem with the race occurs at the starting line. The outcomes in a market economy depend to a considerable extent on the resources that people can bring to the marketplace, and some of these resources are beyond their control. Those born into high-income families with highly educated parents have a much better chance of succeeding than those born into low-income families with poorly educated parents. A person's genetic makeup also matters. Some people are naturally bright, outgoing, and competitive, traits that tend to be rewarded in the marketplace. Others possess special talents such as exceptional athletic ability that are very highly rewarded. Still others lack any of these traits. In effect, then, people are forced to begin the economic race to well-being at very different starting lines through no fault of their own. Given the widely unequal chances of success, many people are quite willing to make independent judgments of the outcomes according to their perceptions of end-results equity and to adjust the outcomes by redistributing if necessary.

Of course, people may be quite willing to judge economic outcomes without much concern about the underlying process that generated them. For example, they may simply take pity on the poor without caring how they became poor. Whatever the motivation, the quest for end-results equity figures prominently in normative public sector theory.

1. Refer to the works of Nozick (1974) and also Hal Varian's excellent mainstream critique of Nozick's position in Varian (1974–1975).

End-Results Equity

End-results equity has proven to be an extremely elusive concept. The quest for end-results equity is often termed the quest for distributive justice, that is, a just distribution of income, but trying to determine the just distribution of income runs into a fundamental difficulty that can be seen in terms of redistributing income toward the "just" distribution. Suppose the government engages in a tax-transfer program in an attempt to reach the just distribution. How large should the program be? To know when to stop redistributing, the government must somehow compare the losses of the losers (those who are taxed) with the gains of the gainers (those who receive the transfers). Unfortunately, no one, not economists or anyone else, has ever come up with a compelling way to do this. Indeed, economists are skeptical of any attempt to make interpersonal comparisons of well-being. Yet some means of comparing gains and losses across people must be made for end-results equity to be operational; otherwise, no one can know how much to redistribute to arrive at a distribution that is "just."

In truth, all we have is a range of suggestions to serve as guidelines for end-results equity. To give one example, Lester Thurow argues that there is a strong bias for equality in the United States, so strong that the burden of proof is on inequality—inequality in the distribution of income always has to be justified (Thurow, 1975). The most common economic justification for tolerating inequality rests on efficiency grounds, that the taxes and transfers used to redistribute generate inefficiencies in the economy. Most economists would argue that the marginal inefficiency costs of further equalizing the distribution outweigh the marginal benefits in terms of end-results equity at a point well short of full equality.

Thurow's position on the bias toward equality may seem extreme, but we will see in Chapter 4 that it has generally been incorporated into public sector theory. The models commonly used by public sector economists to express a concern for end-results equity have the property that everyone should end up with the mean level of income if taxes and transfers do not generate any inefficiencies.

The only widely accepted norm within end-results equity is the principle of *horizontal equity*, which calls for equal treatment of equals: Two people who are equal in all relevant economic dimensions, such as ability and productivity, should enjoy an equal amount of well-being. We will see that horizontal equity has considerable standing among public-sector economists in the design of tax policy. Horizontal equity also provides a link between process equity and end-results equity. Equal opportunity in the marketplace leads to horizontal equity; equal treatment of equals is a requirement for a long-run equilibrium if markets are competitive, with no barriers to entry and exit.

A related principle of end-results equity is *vertical equity*, which says that unequals may be treated unequally. This principle, even if accepted, begs the difficult question of just how unequally society should treat unequals. We know that people who are unequal in ability and productivity can be treated very unequally in a market economy, even if markets are perfectly competitive. Some earn fabulously high incomes, while others do not earn enough to escape poverty. How much inequality should be tolerated? There is no consensus at all on this question, which is hardly surprising. After all, the quest for vertical equity is the same as the quest for distributive justice.

The Government as Agent

The humanistic value judgment of consumer sovereignty has one final and rather remarkable implication for normative public sector theory that concerns the way the government should proceed in designing its policies. The government is not supposed to have a will of its own, in the sense that government officials are not permitted to interject their own preferences into the design of policy. Instead, the proper role of the government is that of an agent acting on behalf of the citizens. The idea is this. Suppose that the market system fails in some way that legitimizes government intervention. The government is expected to design policies to set the economy back on the path toward efficiency or equity, but in doing so it should follow only the preferences of its citizens. The preferences of the president or the members of the legislature carry no special weight; these people are just one of the many citizens with one voice and one vote. Their only job is to accurately represent the desires of their constituents.

The government-as-agent viewpoint has considerable standing in the United States. It is essentially the view expressed by Abraham Lincoln in his Gettysburg Address when he referred to the government being of the people, by the people, and for the people. Lincoln was simply reminding us that the purpose of democratic or representative forms of government is to follow the will of the people. Nonetheless, accepting this view of government severely limits the scope of public sector theory. It implies that the theory is not meant to be a theory of government behavior in the sense of recognizing the state as an organic being with a (political) life of its own. It also consciously removes the theory from the reality that government officials often interject their own preferences into the decision-making process. They do not simply follow the preferences of their constituents.

Ignoring the preferences of public officials is clearly a severe limitation for a political theory of the government, but it happens to be a source of richness and subtlety for an economic theory. A normative economic analysis based solely on the preferences of some group of government administrators would be little more than an exercise in the theory of consumer behavior: What are the administrators' objectives? What choices are available to them? What constraints are they operating under? These may be interesting practical questions, but they do not carry much normative weight.

By forcing the government to consider only the preferences of its citizens, however, all sorts of interesting and difficult problems arise. For example, what should the government do if individual preferences clash, as they inevitably will? Suppose one group of citizens wants more spending on national defense, while another group wants less spending. How should the government resolve this conflict? Normative theory must provide answers to questions such as these.

Other puzzling questions arise as well about the appropriateness of government intervention. If the market system cannot solve a particular problem, acting as it does on individual preferences, why should the government be able to do any better, if all it has to work with are the same individual preferences? A strict libertarian economist might insist that government intervention can only be justified if markets fail *and* if it can be demonstrated conclusively that some *viable* government policy will actually improve upon the market results. Most economists have been content to assign to normative theory the lesser task of describing a *potential* improvement through government action. But this does leave open the question of whether some normative policy prescription really is viable, and, if not, whether a different, viable, policy can actually improve social welfare.

This question lies at the heart of *social decision theory*, a rapidly expanding subspecialty within public sector economics. Social decision theory analyzes the problem of designing practical decision rules and procedures that can achieve optimal normative policies. One of its main concerns is whether democratic voting procedures are consistent with economic efficiency and equity. Another concern is whether government policies can be decentralized. Suppose a market goes astray for some reason and generates nonoptimal outcomes. The preferred solution is to let the market continue to operate but nudge it with policies toward the optimal outcome. This solution is decentralized in the sense that the individuals and firms remain the decision makers in the market. The alternative to decentralization is government provision or some form of coercion. This may be inevitable to solve some problems, but it is never the preferred choice.

As one might expect, sometimes there are clear answers to practical questions such as these, and sometimes not. In any event, it is the principle of consumer sovereignty and the government-as-agent perspective that makes them all so compelling.

GOVERNMENT EXPENDITURE THEORY AND MARKET FAILURE

The Fundamental Theorems of Welfare Economics

Since legitimacy for government intervention is defined in terms of market failure, the natural question to ask is "In what sense do markets fail?" To determine the answer, let us begin with the problem of achieving an efficient allocation of resources.

The market system is entirely neutral with respect to society's well-being, of course. Nonetheless, if conditions are right, competitive markets generate an efficient allocation of resources. The problem for a market economy is that the conditions or assumptions underlying a perfectly functioning market system are far too strong. They typically do not hold in practice, and when they do not a public policy can be described that is pareto superior to the free-market allocation of resources. That is, the public policy can reallocate resources so as to make at least one consumer better off without making any other consumer worse off. This principle underlies all normative policy prescriptions concerned with the allocation of resources.

To determine the subject matter of normative public sector theory, then, consider the assumptions that would allow a market economy to achieve a pareto-optimal allocation of resources. These "best" assumptions fall into two distinct groups: a set of *market assumptions* about the structure of individual markets within the market economy and a set of *technical assumptions* about consumers' preferences and production technologies.

The market assumptions are necessary to assure that all markets are perfectly competitive, so that each economic agent is a price taker and acts on full information. This is the case if four assumptions hold:

1. There are large numbers of buyers and sellers in each market.
2. There is no product differentiation within each market.
3. All buyers and sellers in each market have access to all relevant market information.
4. There are no barriers to entry or exit in markets.

The technical assumptions are required to assure that both consumption and production activities are "well behaved," so that perfectly competitive markets do generate a pareto-optimal allocation of resources. Consider the following set of technical assumptions:

1. Preferences are convex.
2. Consumption possibilities form a convex set.
3. No consumer is satiated.
4. Some consumer is not satiated.
5. Preferences are continuous.
6. Individual utility is a function of one's own consumption and own factor supplies.
7. An individual firm's production possibilities depend only upon its own inputs and outputs.
8. Aggregate production possibilities are convex.

Assumptions 6 and 7 rule out the possibility of externalities in either consumption or production. Assumptions 1, 2, and 5 on individual preferences are satisfied by the standard assumptions of consumer theory, that utility functions are quasi-concave, continuous, and twice differentiable. Assumptions 3 and 4 are commonly employed in economic analysis. Assumption 8 on aggregate production possibilities implies constant or increasing opportunity costs and is satisfied if all individual firms' production functions are continuous, twice differentiable, and exhibit either decreasing or constant returns to scale. Assumption 8 rules out significant increasing returns to scale production, which would imply decreasing opportunity costs, or a production-possibilities frontier convex to the origin.

Gerard Debreu has shown that (Debreu, 1959)

1. If assumptions 1, 2, 3, 6, and 7 hold, then a competitive equilibrium is a pareto optimum.
2. If assumptions 1, 2, 4, 5, 6, 7, and 8 hold, then a pareto optimum can be achieved by a competitive equilibrium with the appropriate distribution of income.

Results (1) and (2) are the *two fundamental theorems of welfare economics*.

Debreu's fundamental theorems of welfare economics have the following implication for public policy. If the four market assumptions hold so that all markets are perfectly competitive, and the combination of technical assumptions specified under (1) or (2) of the fundamental theorems of welfare economics hold as well, then the government sector would not be required to make any decisions regarding the allocation of resources. Indeed, it would not be permitted to do so, according to the normative ground rules. Everything would be left to the marketplace.

The Distribution of Income

If all the appropriate market and technical assumptions hold, would there be anything at all for the government to do? The answer is yes, because of society's concern for end-results equity. A perfectly functioning market system can assure an efficient allocation of resources. Perfect competition also satisfies the process equity norm of equality of opportunity and is likely to generate a high degree of social mobility. But, even a perfectly functioning market economy cannot guarantee that the distribution of the goods and services will be socially acceptable. As noted above, the market takes the ownership of resources as a given at any point in time. If society deems the pattern of

ownership to be unjust, then it will probably find the distribution of goods and services produced by these resources to be unjust as well. Moreover, there are no natural market mechanisms to correct for distribution imbalances should they occur, nothing analogous to the laws of supply and demand, which, under the stringent conditions listed above, automatically select pareto-optimal allocations. Thus, a decision concerning the distribution of income is the first order of business in public sector economics in the sense that it cannot be assumed away. Even in the best of all worlds, with all the appropriate market and technical assumptions holding, the government has to formulate some policy with respect to the distribution of income if society cares about end-results equity. Society might simply choose to accept the market-determined distribution, but this is still a distribution policy requiring a collective decision on the part of the citizens even though it involves no actual redistribution. Moreover, no country has ever made this choice. At a minimum, then, a normative theory of the public sector must address the fundamental question of distributive justice: What is the optimal or just distribution of income?

We have already noted that the search for an optimal income distribution has not achieved a consensus. The only point to add is that any attempt to solve the distribution question is at odds with the preferred government-as-agent ground rule that follows from the principle of consumer sovereignty. By its very nature, a redistribution of income must violate the principle of consumer sovereignty, so long as the losers in the redistribution do not willingly surrender some of their incomes. Therefore, redistribution policy cannot be based entirely on consumers' preferences, with the government simply acting as a passive agent responding to their preferences. It requires a collective decision articulated through some kind of political process, one in which government officials are likely to play a very active role. Normative public sector theory cannot be entirely devoid of political content. Politics necessarily enters the theory through society's attempt to resolve the distribution question.

The collective political decision is troublesome for normative public sector theory, however, because of the lack of a consensus on a set of distribution norms to guide the decision. Furthermore, the theoretical difficulties spread far beyond the distribution question. Since an economic system is a closed system in which all decisions are ultimately interrelated, any public policy decision on the distribution of income necessarily affects all the allocational issues as well. The government cannot simply make a particular redistributional decision, for better or worse, and be done with it.

Public sector economics has never totally come to grips with this problem. Economists have all too often assumed away distributional problems in order to analyze more comfortable allocational issues, knowing full well that separating allocational and distributional decisions is often not legitimate and may produce normative policy prescriptions quite wide of the mark. Some theoretical studies that do incorporate distributional considerations into their models make no attempt to justify particular distributional norms. Rather, the government's distributional preferences are simply taken as given, and normative policies are described with respect to these preferences. The spirit of the analysis is to "have the government provide us with a set of distributional preferences, and we will tell it what it should do." Perhaps this is all economists can hope to do with the distribution question, but it is at least unsettling that the resulting policy decision rules depend upon an assumed pattern of distributional preferences that has no special normative significance.

The Allocation of Resources

The allocational issues in public sector economics follow directly from a breakdown in the market and technical assumptions necessary for a perfectly functioning market system. Many of the market and technical assumptions do fail to hold in practice, so there is broad scope for legitimate government activity. A long tradition within the profession held that the study of failures in the market assumptions typically fell within the domain of industrial organization or consumer economics. These fields analyze such problems as monopolistic behavior and imperfect information, along with the corresponding public policy responses such as antitrust and consumer-protection legislation. Public sector economics, or public finance, traditionally limited its concern to breakdowns in the technical assumptions,[2] concentrating primarily on *externalities* and *increasing returns* or *decreasing cost production*.

Private or Asymmetric Information

This traditional division has broken down in one respect over the past 40 years, around the problem of imperfect

2. The theory of fiscal policy can also be thought of as a response to a breakdown in the market and technical assumptions. For example, externalities play a role in the two main themes of macroeconomic policy, stabilizing the business cycle and promoting optimal long-run economic growth. New Keynesians argue that coordination problems are an important determinant of the wage and price stickiness that gives rise to the business cycle from the demand side. The economy would operate closer to its production frontier, on average, if workers and firms would agree to index wages and prices to the rate of growth in aggregate demand. But individual firms and workers are not willing to index unless they can be assured that all workers and firms will index, and coordinating an economy-wide indexing is difficult to accomplish in practice. Therefore, wages and prices remain largely unindexed. Similarly, externality problems help to explain why a nation's rate of saving might not be optimal, at a rate consistent with the Golden Rule of Accumulation, which maximizes consumption per person over time. Externalities are also central to the newer endogenous theories of long-run economic growth (for instance, all those theories that point to the spread of knowledge as an engine of growth).

information. Economists have been particularly interested in the consequences of asymmetric information, in which some individuals have private information that other individuals do not know. Private or asymmetric information is so common in exchange that it has become a focus of analysis in all fields of economics, including public sector economics. Some reflection on the relationship of private information to government policy is in order, because economists have come to realize that private information has a profound effect on normative public sector theory.

Private information is, first of all, an important source of market failure that requires government intervention. The general problem with private information is that it tends to undermine market exchanges because it gives an undue advantage to those who have it. They can easily cheat the other parties. This is why even the most libertarian of economists acknowledges the need for a judicial system to enforce contracts and define private property rights. It also leads to agencies such as a bureau of standards to protect consumers from fraud (e.g., to ensure that a gallon of gasoline at the pump really is a gallon), and the Occupational Safety and Health Administration to ensure that workers understand the hazards of their jobs. People want independent certification from the government that producers are telling the truth about products and working conditions.

The widespread provision of public insurance is another important example of a response to market failure caused in part by private information. Private firms are willing to provide insurance against risky events only if a number of conditions hold. Among them is the requirement that they have good information about the insured. Absent good information, the insurance companies are exposed to the *principal—agent problem.* The structure of the problem is that a principal is in charge of a set of agents who have different objectives from the principal. Therefore, the principal has to monitor the agents so that they will behave in accordance with the principal's objectives, and the principal needs good information about the agents to monitor them effectively.

In the case of insurance markets, each insurance company (the principal) needs to be able to monitor the insured (the agents) to write profitable policies. For starters, the companies need to know the riskiness of the insured so that they can adjust their premiums according to risk (e.g., higher auto insurance premiums for the more risky drivers). Otherwise, they are forced to charge one premium for all risk classes, and the low-risk policy holders have an incentive to drop out and form their own group. This phenomenon is called *adverse selection*, because it leaves the insurance companies with an ever-riskier (adverse) pool of the insured, and the companies must charge ever higher premiums to earn a profit. At some point, the premiums may become too high to attract a large enough pool of high-risk policy holders, leaving the high-risk people without any insurance. Insurance companies also have to be confident that their policy holders cannot influence the probability of the event being insured against unbeknownst to the company (e.g., unhealthy lifestyles that are difficult for the medical insurers to detect). The ability to change the odds of the insured event is called *moral hazard*, and it is a clear threat to the profitability of the insurance companies. Private firms may not provide insurance if either adverse selection or moral hazard is a possibility; consequently, people who want the insurance must turn to the government to provide it. In fact, the governments in most of the developed market economies operate large public insurance programs.

At a deeper level, private information threatens the government-as-agent role that the government is supposed to play when trying to solve allocational problems. The government obviously must know the preferences of the people to be an effective agent on their behalf. But if people have private information, they often have an incentive to hide their true preferences from the government to get a better deal for themselves by having others "play the sucker." The government cannot hope to achieve pareto-optimal allocations if the people will not reveal their preferences, as pareto optimality is defined in terms of each individual's own preferences.

Unfortunately, getting self-interested people to tell the truth is a difficult problem in the context of many allocation issues, as we shall see throughout the text. A major research agenda in social decision theory is the *mechanism design problem*: how to design preference-revealing mechanisms such that the dominant, utility-maximizing strategy is for people to reveal their true preferences. Some truth-revealing mechanisms have been described, but most are not practicable. The one exception has been in the design of auctions used by the federal government to sell rights to oil reserves and telecommunication bandwidths.

Getting people to reveal the truth about themselves is also a central problem in designing tax and transfer policies. Governments do not want people to escape taxes or receive inappropriate transfers by claiming to be something other than what they are. Economists have been successful in designing tax-transfer policies that are truth revealing, but having to design the policies in this way still undermines the government-as-agent ideal because it wastes resources relative to the case of perfect information. (See later discussion of tax theory.)

At the deepest level, private information can be viewed as the fundamental justification for *all* government intervention directed at allocational problems. To see why, suppose that everyone did have full information, as Debreu's fundamental theorems of welfare economics assume. If so, then self-interested individuals would presumably use their knowledge to extract all possible

pareto-superior gains from the economy because they have a mutual interest in doing so. They would employ whatever means are necessary—markets, various forms of private negotiation and bargaining, and side payments to exploit all the gain−gain opportunities. The economy would naturally achieve a pareto-optimal allocation of resources, without the aid of any kind of government policy. This would be true even if the other market and technical assumptions failed to hold. The economy could be riddled with market power, externalities, and decreasing cost production. Yet self-interested agents with perfect information would discover the pareto-superior allocations for all these problems.

The only limitation on these private exchanges would be the transaction costs of making them, which Debreu's analysis assumed away. The transaction costs might exceed the potential gains from an exchange in some cases, but to argue that transaction costs are a justification for government intervention under perfect information is not entirely convincing. People are unlikely to have perfect information about each other if significant transaction costs hinder their exchanges and negotiations. The assumptions of perfect information and insignificant transaction costs tend to go hand in hand. Furthermore, if transaction costs prevent private exchanges from occurring they may also prevent government agencies from improving on the private allocations. Why should the government have an advantage in reducing transaction costs over coalitions of private citizens armed with perfect information?

The only obvious role for the government under perfect information would be distributional, to redistribute income if necessary in accordance with society's norms regarding end-results equity. There would be no need for any normative economic analysis relating to allocational problems, not in public sector economics or in any other field of economics. Therefore, private information may well be the ultimate justification for government intervention in correcting all allocational inefficiencies.

THE GOVERNMENT SECTOR IN THE UNITED STATES

Limiting the allocational functions of government to externalities, decreasing cost production, and private information within public sector economics may seem highly restrictive, yet nearly all the exhaustive or resource-using expenditures on goods and services in the United States can be justified in terms of these conditions. We have already noted the justification of the judicial system, various bureaus of standards or safety, and public insurance programs on the basis of private information. Examples of US government programs justified in terms of externalities include defense, the space program, and related activities, which together comprise the overwhelming majority of exhaustive expenditures in the national budget; education, which accounts for nearly 40% of all state and local exhaustive expenditures; and many lesser items such as local public safety and government-supported research and development programs. Public services exhibiting significant increasing returns-to-scale production include many types of public transportation (which frequently generate externalities as well), the public utilities (electricity, water, and sewerage), many recreational facilities (public parks and beaches), and radio, television, and other forms of communication such as the Internet, which may well be among the purest examples of decreasing cost services.

Table 1.1 lists the expenditures of the US federal, state, and local governments for fiscal year (FY) 2012 (federal) and FY 2010 (state and local), the last years that the data were available as this was written. The data underscore the view put forth in this introductory chapter that market failure is the primary justification for government intervention in the United States. On the one hand, most of the resource-using purchases of goods and services exhibit either externalities or increasing returns. On the other hand, purchases of goods and services accounted for only 22% of total federal expenditures in FY 2012. The remainder were transfer payments: transfers to persons or grants-in-aid to state and local governments or interest payments on the national debt. The transfers to persons, the largest category, are primarily redistributive in their impact.[3] As such, they too can be considered a response to market failure, namely, the inability of the market system to guarantee an acceptable distribution of income. Also, a large proportion of the grants-in-aid help the state and local governments pay for two of the largest public assistance programs targeted to the poor, Temporary Assistance to Needy Families and Medicaid. These two programs are administered by the states (and localities in some states). Finally, the largest single government program, Social Security (including Medicare), reflects a mixture of motives based on market failure: redistributional (the elderly are vulnerable to becoming impoverished in a market economy without public pensions); insurance (relating to uncertainty about the timing of death and the problems of private information inherent in medical insurance); and paternalism (without the forced savings through payroll taxes to pay for Social Security benefits, many people might not save enough for their retirement and would risk becoming wards of the state).

3. As noted above, the large public insurance programs have an informational justification. Nonetheless, the problems of adverse selection and moral hazard do not disappear with government provision of insurance. Public insurance programs inevitably redistribute from low-risk to high-risk individuals and from the honest to those engaging in moral hazard. These unintended redistributions may help to explain why public insurance programs are strenuously opposed by so many taxpayers.

TABLE 1.1 Expenditures by Federal, State, and Local Governments in the United States

	Expenditures ($, Billions)	Percentage of Subcategory	Expenditures ($, Billions)	Percentage of Total Expenditures (%)
A. Federal Government (FY 2012)				
Government expenditures on goods and services			788	22
Defense and defense related[a]	688	87		
Nondefense	100	13		
Domestic transfers to persons (direct expenditures)			1886	53
Social insurance and pensions				
Social security benefits (Old Age Survivors and Disability Insurance-OASDI)	773	41		
Medicare	555	29		
Civilian and military retirement	129	7		
Unemployment insurance	96	5		
Agricultural support payments	10	1		
Veterans benefits[b]	124	7		
Student assistance	44	2		
Public assistance				
Supplemental Nutrition Assistance Program-SNAP (food stamps)	80	4		
Housing assistance	51	3		
Supplemental security income (SSI)	51	3		
Earned income tax credit (EITC)	55	3		
Net interest payments			232	7
Grants-in-aid			632	18
Payments to individuals	399	63		
Temporary Assistance for Needy Families-TANF	16	6		
Medicaid	251	63		
Other	233	37		
Total expenditures			3538	100
B. State Governments (FY 2010)[c]				
Direct expenditures			1108	70
Public welfare	404	36		
Education	254	23		
Highways	93	8		
Health and hospitals	99	9		
Other	258	23		
Grants-in-aid			486	30
Total general expenditures			1594	100

Continued

TABLE 1.1 Expenditures by Federal, State, and Local Governments in the United States—cont'd

	Expenditures ($, Billions)	Percentage of Subcategory	Expenditures ($, Billions)	Percentage of Total Expenditures (%)
C. Local Governments (FY 2010)				
Education			605	42
Housing and community development			42	3
Health and hospitals			126	9
Public safety			154	11
Public welfare			52	4
Highways, airports, other transportation			115	8
Other			336	23
Total general expenditures			703	100

[a]*Includes national defense; general science, space, and technology; and international affairs.*
[b]*Includes education benefits, medical benefits, insurance benefits, and compensation, pension, and burial payments.*
[c]*Data for state and local governments were available only through fiscal year 2010.*
Sources: U.S. Department of the Treasury, "Monthly Treasury Statement," September 2012, www.fms.treasury.gov/mts0912.pdf. U.S. Census Bureau, J. Barnett and P. Vidal, "State and Local Government Finances Summary: 2010," Appendix Table A1, Government Division Briefs, September 2012, www2.census.gov/govs/local/10_summaryreport.pdf.

THE THEORY OF TAXATION

Most of the remarks thus far have been directed to the theory of public expenditures as opposed to the theory of taxation, because the former is logically prior to the latter. Public expenditure theory defines the legitimate areas of public concern as well as the permissible forms that policy may take. Moreover, as indicated above, public expenditure theory often contains its own theory of taxation in the sense that the expenditure decision rules define a set of taxes and transfers necessary to guide the market system to an optimum. Taxes contribute to the pursuit of efficiency and equity in these instances.

The theory of taxation becomes interesting in its own right only when the expenditure decision rules indicate the need for specific government expenditures without simultaneously specifying how those expenditures are to be financed. When this occurs, the same criteria that guide public expenditure analysis also apply to the collection of tax revenues. In particular, taxes should promote society's microeconomic goals of allocational efficiency and distributional equity.

A natural tension arises between tax policy and the goal of allocational efficiency, however. Most taxes generate distortions in the market system by forcing suppliers and demanders to face different prices. These distortions misallocate resources, thereby generating allocational inefficiencies. Resource misallocation is not desirable, of course, but it is an unavoidable cost of having to raise tax revenues. One goal of normative tax theory, then, is to design taxes that minimize these distortions for any given amount of revenue to be collected. Alternatively, if the government must use one of two or three specific kinds of taxes to raise revenue, normative tax theory should indicate which of these taxes generates the minimum amount of inefficiency.

Normative issues such as these are part of the allocational theory of taxation and, just as with the allocational issues of public expenditure theory, the guiding principle is pareto optimality. According to the pareto criterion, the government should collect a given amount of revenue such that it could not raise the same amount of revenue with an alternative set of taxes that would improve at least one consumer's welfare without simultaneously lowering the welfare of any other consumer. If such pareto improvements are impossible, then tax policy satisfies the pareto criterion of allocational efficiency, even though it necessarily generates inefficiencies relative to a no-tax situation.

The second unavoidable effect of taxes is that they reduce taxpayers' purchasing power so that they necessarily become part of the government's redistributional program. The government naturally wants its taxes to contribute to society's distributional goals, but there are two difficulties here. The first is that the redistributional theory of taxation suffers from all the indeterminacies of redistributional theory in general. Thus, while public sector economists generally agree on normative tax policy with respect to society's allocational goals, there is considerable disagreement as to what constitutes good tax policy in a distributional sense. The second difficulty is the inherent trade-off between equity and efficiency in taxation. Generally speaking,

achieving greater redistribution requires levying higher tax rates on the "rich" but, as we shall discover, higher tax rates tend to increase inefficiency. In addition, taxing a particular good might be desirable in terms of society's distributional goals but highly undesirable on efficiency grounds, or vice versa. Understanding the nature of these kinds of equity−efficiency trade-offs has always been a primary goal of normative tax theory.

Two additional subsidiary goals of tax policy are *ease of administration* and *simplicity*, which relate to the practical problem of collecting taxes. The ease of administration criterion adopts the tax collectors' point of view. A tax has to be easy for a department of revenue to administer or it will not be used. Private information comes directly into play here. Self-interested taxpayers have a strong incentive to avoid paying taxes, and they can do so if they are able to hide information about themselves from the government's tax collectors. Illegal avoidance of taxes is called *tax evasion.* Legal sanctions or just plain old honesty may prevent some people from cheating on their taxes, but not everyone. Therefore, the design of any tax has to address the problem of potential evasion.

Consider an income tax as an example. Suppose the government wants to tax high-income taxpayers at a higher rate than low-income taxpayers as part of its redistributional policy. It may not be able to do this, however, if high-income taxpayers can hide much of their income from the authorities and thereby evade much of their proper tax liability. Also, the hiding of income forces the government to raise average tax rates to collect a given amount of revenue, which increases the inefficiencies associated with the tax. Finally, some taxes are easier to evade than others. Therefore, the relative ease of evading different taxes has to be considered in determining what mix of taxes to use to meet the government's total revenue requirements.

The goal of simplicity adopts the taxpayers' point of view. Taxpayers have to be able to comply with the tax laws fairly easily for a tax to be used. They must be able to understand the tax laws and not suffer undue recordkeeping and filing burdens. A clear example of this principle is the preference in less-developed countries for taxing businesses rather than people. The average person is not educated enough to maintain records on income or prepare and file an income tax form, regardless of how honest or dishonest he or she may be. Therefore, the less-developed countries tax businesses simply because they are able to collect taxes on businesses.

FISCAL FEDERALISM

A hierarchical structure of national, state (provincial), and local governments raises a number of interesting normative issues that cannot arise with a single government. Foremost among them is the question: What is the advantage of having layers of governments as opposed to a single national government? In terms of the prevailing jargon, should government be decentralized or centralized? The conventional wisdom within democratic societies is that a highly decentralized federalism is preferable because local government officials know the preferences of their citizens better than national officials do. Therefore, each legitimate function of government should be provided at the lowest level of government in the fiscal hierarchy, consistent with the requirements of efficiency and equity.

Counterbalancing this conventional wisdom are some difficult problems associated with the ability of people to move from locality to locality in response to local government policies. The ability to move can itself generate inefficiencies that would not be possible with a single government. It also raises the possibility of multiple equilibria or no equilibrium at all as people search for the localities that maximize their utilities. Mobility also severely limits the possibilities for redistributing income at any level in the fiscal hierarchy other than the national level. Suppose a locality undertakes a tax-transfer policy to redistribute income from its high-income citizens to its low-income citizens. The high-income citizens have an incentive to move to another locality that is not redistributing, thereby undermining the original locality's redistribution policy and lowering the average income in the locality as well. At the same time, we shall see that denying a government the distribution function removes its political identity in the mainstream model of the public sector. This leads to another fundamental problem for a normative theory. With each person simultaneously being a citizen of multiple governments and with some of the governments lacking political identities, the notion of an overall social optimum that the various governments are striving for becomes highly problematic.

Information also plays a special role in the normative theory of fiscal federalism. The main issue here is how sophisticated people are within each local government. As they vote on policies in their own localities, do they consider how people in other localities might react to their policies, or do they take the policies elsewhere as given? The answer to this question has important implications for the efficiency of local solutions to allocational problems.

THE THEORY OF PUBLIC CHOICE

The theory of public choice developed by James Buchanan and his followers challenges virtually every tenet of the mainstream public sector theory. Buchanan described the foundations of the public choice perspective in his Nobel lecture delivered in Stockholm, Sweden, in 1986.[4] The

4. The lecture was reprinted in Buchanan (1987).

disagreements with the mainstream view begin at the most basic level, with the assumptions about how people behave. According to Buchanan, the mainstream theory assumes that people are essentially schizophrenic. They are self-interested in their economic lives, but when they turn to the government in their political lives they suddenly become other-interested and consider the broader social or public interest in efficiency and equity. Nonsense, say the public choice advocates. People do not change their stripes; they remain self-interested in their political lives as well. They turn to government only because they cannot get what they want for themselves in the marketplace, and they view the government as just another venue for seeking their own objectives. Buchanan refers to individuals' interactions with the government as fiscal exchanges, to mirror the self-interested motivations of standard market exchanges. Using the government in the pursuit of self-interest is seen as entirely appropriate and legitimate.

The thrust of public choice theory is positive, not normative. Buchanan scoffs at the notion of an idealized, beneficent government acting as an agent of the people in pursuit of social objectives. Instead, Buchanan argues that public sector economists should be studying actual political and governmental institutions and determining whether they give the people what they want. The test of government efficiency in this positive vein is simply how well the government serves each person's self-interest. Full efficiency requires unanimity under democratic decision making, because only then will no one lose as a result of any government policy. This is as "efficient" as the government can be in helping people get what they want. Notice that the public choice definition of efficiency in political activity is far stronger than the economic definition of efficiency as pareto optimality, which the mainstream perspective uses to judge public policies.

The public choice perspective does have normative content but it is strictly process oriented, concerned only with the rules that govern political activity. Moreover, Buchanan claims that the normative content centers on a single point in time, at the founding of a democratic nation. The norms are embedded in the constitution drafted by the nation's constitutional convention.

In focusing on the constitution, Buchanan was influenced by the Swedish economist Knut Wicksell, who theorized about the legitimate role of government in a democratic society at the end of the nineteenth century. It was Wicksell who first thought of government activity in terms of fiscal exchanges and who described the ideal as unanimous consent for all policies at every point in time. Buchanan concedes that requiring unanimity all the time is asking for too much; it would lead to paralysis. Instead, he points to the constitution. He argues that legitimacy in government requires only a consensus among the framers of the nation's constitution about the rules under which the government is permitted to operate. In designing these rules, the convention members think only of their self-interests and those of their descendants as they perceive them. Unanimous agreement at the constitutional convention about the rules of politics would be the ideal, although Buchanan concedes that a consensus may be all that is possible.

The only valid normative test of government activity at any time after the convention is the following: Could the current rules that guide and constrain government activity have arisen from an agreement at the constitutional convention? If the answer is yes, then the current rules are legitimate and society has forged a legitimate link between the people and their government. Notice that the policies that result from these rules cannot be evaluated directly by any norms. In particular, the outcomes of policies are irrelevant in and of themselves. Process is everything according to this test, namely, consistency with the self-interested rules agreed to at the constitutional convention.

Normative policy analysis after the convention is possible, but it is limited to suggestions for constitutional reform and then only if the normative test fails. Normative proposals take the form of recommending changes in the constitutional rules so that people are better able to pursue their self-interests in their fiscal exchanges with the government. For example, Buchanan seriously doubts that the large, prolonged US federal budget deficits that have existed in most years since the early 1980s would pass his normative constitutional test because of the damage they could inflict upon future generations. He favors a balanced-budget amendment to the constitution.

An interesting question is whether redistributional policies or rules could ever achieve a consensus at a constitutional convention, given that redistributions force some people to pay taxes for the benefit of others. Those who are taxed may well feel that they are not getting what they want from their fiscal exchanges. Buchanan believes that consensus could be reached if the framers of the constitution choose to consider the welfare of future generations and are willing to view the future through a veil of ignorance. The idea is that no one can predict the future, so that no one at a constitutional convention can know with certainty how their descendants will fare for all time. Therefore, they may see it in their self-interest to establish rules that permit redistributions of income on the chance that their descendants might be the ones who fall on hard times. In other words, they are simply allowing for the possibility of future transfers to their own families.

The public choice perspective is persuasive in a number of respects. The assumption of self-interested political behavior is instinctively appealing to economists, and much political behavior is clearly self-interested. The insistence

on analyzing actual political institutions and actual political choices is also sensible, as is a focus on the constitutional rules that guide and constrain all political activity. Nonetheless, public choice has not captured the day among public sector economists. It remains a distinctly minority perspective, if the weight of the professional literature is an accurate guide.

Perhaps the mainstream has stood firm against the public choice challenge because the normative basis of public choice theory is so thin. The public choice perspective as articulated by Buchanan lacks any clear sense of good citizenship or empathy, qualities that many people believe are essential ingredients for a society that anyone would want to live in. A narrow focus on self-interested constitutional rules may not be enough to sustain a comprehensive normative economic theory of the public sector. In any event, the majority of economists apparently want to judge the results of specific government policies directly and to do so in terms of the pareto efficiency criterion and commonly accepted equity norms such as equal opportunity or horizontal equity. More generally, government activity motivated entirely by self-interest simply does not have the normative appeal of government activity motivated by the public interest in efficiency and equity.

The battle between public choice and mainstream economists is unlikely to be decided on empirical grounds because ample evidence exists to support both sides. Two published reflections by Joseph Stiglitz and Joel Slemrod are instructive.[5]

Stiglitz, a Nobel laureate, has contributed as much as any economist to mainstream public sector theory over the past 50 years. When he was asked to reflect on his years at the Council of Economic Advisors, he responded with a paper describing why the government has such difficulty enacting policies that are so clearly beneficial from the mainstream perspective. The problem in a nutshell, according to Stiglitz, is that all too many government officials behave as Buchanan said they would. They pursue and protect their self-interests rather than the public interest, such as by keeping their private information secret when it is to their personal advantage to do so. Stiglitz believes that the government is hugely beneficial overall but not nearly so much as it could be if officials were more consistently public spirited.

Joel Slemrod has been a major contributor to mainstream tax theory and policy over the past 35 years. He recently speculated that other-directed, civic-minded behavior may produce much more than just a kinder and gentler society. He points to some studies that show a positive relationship between economic growth and prosperity and what he terms social capital, such things as the degree of trust in others, the propensity to obey society's rules, and civic behavior. The social capital variables in these studies are obtained though surveys. A connection between civic-minded, other-directed behavior and economic growth would be a major boost for the mainstream perspective if it stands up to further analysis.

BEHAVIORAL PUBLIC FINANCE

Starting around 1970, economists began to consult and work with psychologists and psychiatrists to better understand the nature of preferences. As that line of research grew it uncovered all kinds of behavior that was anomalous from the perspective of mainstream economic theory because it was inconsistent with rational self-interest. At the same time, however, the behavior could be seen as consistent with basic psychological principles of behavior. Two common examples of these anomalies are (1) framing effects, that people will make different decisions in a given situation depending on how the situation is presented to them, e.g., employee participation in pension plans increases dramatically if the default option is participation rather than nonparticipation; and (2) people exhibit present-biased preferences, also called self-control problems, e.g., smokers and drinkers often know it is in their best long-run interests to quit but do not have the will power to do so.

The study of these anomalies became known as behavioral economics, and branches developed along each of the standard fields within economics, one such branch being behavioral public finance. A main line of research within behavioral public finance is positive in nature: how can policy makers exploit anomalous behavior such that policies are better able to meet their intended goals. The idea is that a better understanding of psychological principles will lead to more effective economic policies. The normative implications of anomalous behavior cut more deeply and are highly controversial among mainstream economists. That people might not always attempt to maximize their own self-interest violates the fundamental assumption of mainstream economic theory, and along with it the mainstream economic theory of the public sector presented in this textbook. Mainstream economists, while conceding that many of the anomalies uncovered by the behavioral economists are widespread and important, are understandably reluctant to give up the many advantages of the mainstream theory. In addition, the behavioral economists have not been able to develop a comprehensive psychologically based theory of behavior to replace the mainstream theory. Chapter 25 explores the behavioral anomalies that are especially important to public sector theory, along with a selection of their positive and normative implications.

5. Refer to the works of Stiglitz (1998), Slemrod (1998), and also a set of lectures by Buchanan and Richard Musgrave, dean of the mainstream economists, recently published in Buchanan and Musgrave (1999).

SUMMARY

To summarize the main points of this wide-ranging overview:

1. Chapter 1 has discussed the predominant themes in the normative economic theory of the public sector as that theory has evolved in Western economic thought. The four foundational elements of the mainstream theory are the following:
   a. Government activity is justified strictly in terms of competitive market failure. In particular, the microeconomic theory of the public sector focuses on the problems caused by externalities, decreasing cost production, asymmetric or private information, and an inequitable distribution of income, none of which can be resolved adequately by the free-market system.
   b. The principle of consumer (and producer) sovereignty is the fundamental value judgment underlying normative public sector theory, that consumers (and producers) are the best judges of their self-interest and should be allowed to pursue their self-interest. Consumer sovereignty ties public sector theory closely to the free-market system, as advocates of market capitalism also embrace the principle of consumer sovereignty.
   c. Government policies should promote the microeconomic goals of allocational efficiency and distributional equity. Allocational efficiency is pareto optimality defined in terms of individuals. Distributional equity includes both process equity and end-results equity. Two widely held norms within process equity are equal opportunity and social mobility. There are no widely held norms within end-results equity other than horizontal equity, which says that equals should be treated equally. Horizontal equity is the one bridge between process equity and end-results equity because equal opportunity generates horizontal equity in the long-run competitive equilibrium. Despite the lack of consensus on other end-result norms, most models used by public sector economists embrace the goal of equality in the sense that inequality has to be justified. The usual justification is the inefficiency of taxing and transferring; at some point, the gains to further equality are offset by the costs of increased inefficiency.
   d. When addressing allocational issues, the government should act as an agent on behalf of the citizens and design policies strictly in accordance with their preferences. The preferences of government officials are irrelevant, other than in their role as citizens. The government-as-agent prescription breaks down if society undertakes redistributional policies in the name of end-results equity. Redistributional policy requires a collective decision through a political process, and it is this collective distributional decision that constitutes the political content of normative public sector theory.
2. Almost all government expenditures in the United States can be justified as reactions to market failures. Most of the exhaustive or resource-using expenditures are reactions to allocational problems resulting from externalities, decreasing costs, and private information. The transfer payments are largely motivated by concerns about the distribution of income, particularly the problem of poverty. The Social Security pensions and Medicare have a mixture of allocational and redistributional motives.
3. The theory of public choice is the primary competitor to the mainstream theory. It assumes that people are motivated in their political behavior by self-interest just as in their economic behavior. The main thrust of public choice theory is positive in nature, to study the operation of actual political institutions and determine if they give people what they want. The normative content of public choice is entirely process oriented. It focuses on the rules under which the government operates as set down in that nation's constitution. The only normative test is whether the current rules that guide and constrain political activity could have emerged from a consensus at the constitutional convention. Normative policy analysis is limited to suggestions for constitutional reforms that will better help people to get what they want. Public choice theory remains a minority position among public sector economists, perhaps because its insistence on strictly self-interested political behavior gives it a fairly thin normative base relative to the mainstream theory.
4. Behavioral economics uses psychological principles to understand behavior that is anomalous from the perspective of mainstream economic theory because it is inconsistent with the fundamental mainstream assumption of rationality—that people act to maximize their self-interest. Behavioral economics is rapidly gaining momentum and appears in all branches of economics, including public sector theory where it is called behavioral public finance. Mainstream economists concede that many of the anomalies are widespread and important, but they believe that the behavioral economists are far from developing a comprehensive psychologically based theory of economic behavior that could replace the standard mainstream theory.

With the mainstream themes in hand, Chapter 2 presents a baseline version of the basic general equilibrium model of an economy that will be used throughout the text to develop normative public sector decision rules. The chapter emphasizes how the efficiency and equity norms described in Chapter 1 are incorporated into the formal model.

REFERENCES

Arrow, K., 1951. Social Choice and Individual Values. Wiley, New York.

Buchanan, J., June 1987. The constitution of economic policy. American Economic Review, 243–250.

Buchanan, J., Musgrave, R., 1999. Public Finance and Public Choice: Two Contrasting Views of the State. MIT Press, Cambridge, MA.

Debreu, G., 1959. The Theory of Value: An Axiomatic Analysis of Economic Equilibrium. Wiley, New York (chapter 6).

Nozick, R., 1974. Anarchy, State, and Utopia. Basic Books, New York.

Slemrod, J., September 1998. On voluntary compliance, voluntary taxes, and social capital. National Tax Journal Vol. 51 (3), 485–491.

Stiglitz, J., Spring 1998. The private uses of public interests: incentives and institutions. Journal of Economic Perspectives Vol. 12 (2), 3–22.

Thurow, L.C., 1975. Generating Inequality: Mechanisms of Distribution in the U.S. Economy. Basic Books, New York (chapter 2), pp. 26–27.

U.S. Department of the Treasury, September 2012. Monthly Treasury Statement.

U.S. Census Bureau, Barnett, J., Vidal, P., September 2012. State and Local Government Finances Summary: 2010. Appendix Table A1, Government Division Briefs.

Varian, H., 1974–1975. Distributive justice, welfare economics, and the theory of fairness. Philosophy and Public Affairs 4.

Chapter 2

A General Equilibrium Model for Public Sector Analysis

Chapter Outline

Chapter 2 develops a baseline analytical model of an economy, variations of which have been used for almost all mainstream public sector analysis.

A model must possess four attributes to be useful as a framework for a normative theory of the public sector. First, it must be a general equilibrium model of the economy. All general equilibrium models describe the three fundamental elements of any economy: (1) the preferences of every consumer, (2) the production technologies, and (3) market clearance for all goods and services and factors of production. A particular model may contain other features as well, but the three fundamentals must be present to have a valid general equilibrium model. Second, the model must be flexible enough to consider a broad spectrum of public sector problems, particularly those associated with externalities, decreasing cost production, asymmetric information, the distribution of income, and various issues in the theory of taxation. Third, the model must be designed to highlight the public interest in efficiency and equity, the two main objectives of normative public sector theory. Finally, the model must be compatible with a market economy, since Western public sector economics assumes that the government operates within the context of a market system.

Paul Samuelson presented a model with exactly these attributes in his 1954 article, "The Pure Theory of Public Expenditure."[1] He happened to use the model to analyze a nonexclusive good such as national defense, which is a particular kind of externality. But Samuelson's model proved to be readily adaptable to the full range of public sector problems, and it quickly became the standard model for virtually all mainstream normative public sector analysis. Indeed, Samuelson's model became the standard normative model used by neoclassical economists in every field of economics. Students will recognize the model in Chapter 2 as the baseline general equilibrium model presented in all intermediate and advanced textbooks on microeconomics.

It is absolutely essential to understand the structure of the Samuelson model and the properties of its solution as a prelude to the study of public sector economics. This is the goal of Chapter 2.

1. Samuelson (1954). The following year Samuelson supplemented the mathematical analysis with a geometric presentation in Samuelson (1955). No articles have had any greater impact on public sector analysis.

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00002-9

A BASELINE GENERAL EQUILIBRIUM MODEL

A general equilibrium model can be specified in terms of quantities or prices. The quantity model is the simpler one because it requires fewer assumptions. It can be thought of as an exercise undertaken by an omniscient social planner who dictates all consumption and production decisions and whose objective is the public interest in efficiency and equity.

The fiction of a social planner can be dropped by specifying the general equilibrium model in terms of prices so that the model describes the operation of a market economy. This requires three sets of assumptions about market behavior and market structure. The first set relates to the objectives of individuals and firms in their market exchanges. The standard assumptions are utility maximization by consumers and profit maximization by firms, but these may not always be appropriate assumptions. For example, consumers and firms may choose other objectives when operating in highly complex and uncertain environments, such as bounded rationality by consumers and profit satisficing by firms. The second set of assumptions relates to the structure of markets: Are they perfectly competitive or something else? The final set relates to the market behavior of the government in its dual role as a consumer of some goods and services and a producer of others. For example, does the government engage in exchange at the market prices or at some other prices that it determines? Whatever the government may do, normative public sector theory always assumes that the government's objective is the public interest in efficiency and equity, just as in the social planner quantity model.

The natural place to begin is with the simpler social planner model specified in terms of quantities. Our baseline model assumes that all the technical assumptions necessary for a well-functioning competitive market system apply, so that we can relate the solution of the model to standard competitive market behavior. This will provide an appropriate analytical foundation for introducing breakdowns in the technical assumptions one at a time in Part II, as we explore public expenditure and tax theory in the context of a competitive market economy. The baseline model is also immediately useful for analyzing the problem of achieving an optimal distribution of income, since the distribution problem exists even if all the technical assumptions hold.

Let's begin, then, with the three fundamental elements of any general equilibrium model: individual preferences, production technologies, and market clearance.

Individual Preferences

As noted in Chapter 1, individuals' preferences are the fundamental demand data for all normative public sector analysis under the government-as-agent ground rule. The individual preferences are defined over all goods and services consumed and all factors supplied. Let there be H individuals (households), G goods and services (hereafter, goods), and F factors. Define:

$$\left(\begin{array}{lll} X_{hg} = \text{the consumption of good g by person h.} & h = 1,\ldots,H \\ & g = 1,\ldots,G \\ V_{hf} = \text{the supply of factor f by person h.} & h = 1,\ldots,H \\ & g = 1,\ldots,G \end{array}\right)$$

and let

$$U^h = U^h\left(X_{hl},\ldots,X_{hG};\ V_{hl},\ldots,V_{hF}\right)$$

or simply

$$U^h = U^h\left(X_{hg};\ V_{hf}\right) \quad h = 1,\ldots,H \tag{2.1}$$

represent the ordinal utility function for person h, assumed to be "well behaved."[2] The functions $U^h(\ )$ represent a complete description of individual preferences for the economy, defined over $H*G$ individual goods consumed and $H*F$ individual factors supplied.

Two points about the specification of factor supplies are worth noting. The first is that individuals are assumed to view factor supplies as bad, a necessary evil for gaining command over goods and services. Therefore, factor supplies enter the utility function with a negative sign. For example, if X is the only good, and L, labor, is the only factor, the utility of person h might be represented as

$$U^h = U^h\left(X_h;\ 24 - L_h\right)$$

where 24 represents the total hours in the day, L_h is the number of hours worked per day, and $(24 - L_h)$ is leisure time, the "good."

2. Utility functions are always assumed to be continuous, strictly quasi-concave, and twice differentiable, with all goods and factors infinitely divisible.

The second point is that our baseline model assumes that the supplies of all factors are variable. Some general equilibrium models assume instead that one or more factors are in fixed supply and treat the fixed factors as separate resource or endowment constraints within the economy. Land is a common example. The fixed factors do not need to enter the utility functions because they are not decision variables for the individuals. They appear only in the market clearance equations and production functions as fixed resources to be allocated among the producers. These resource constraints become a fourth fundamental element of the model. Our assumption that factor supplies are variable is the more realistic one, however, especially for labor and capital (saving).

Production Technologies

Production in a general equilibrium model is completely described by the production technologies that relate inputs of factors to the outputs of goods and services. To remain fairly general at this point, specify a separate production function for each output. Define:

$$\left(\begin{array}{ll} r_{gf} = \text{factor } f \text{ used in the production of good } g. & g = 1,\ldots,G \\ & f = 1,\ldots,F \\ X^g = \text{the aggregate amount of good } g \text{ produced.} & g = 1,\ldots,G \end{array}\right)$$

and let

$$X^g = \phi^g\left(r_{g1},\ldots,r_{gF}\right)$$

or simply

$$X^g = \phi^g\left(r_{gf}\right) g = 1,\ldots,G \qquad (2.2)$$

represent the "well-behaved" production function relating the factor inputs to aggregate production of goods g.[3] The functions ϕ^g ( ) represent a complete description of the economy's production technology, defined over G^*F individual inputs and G aggregate goods and services.

Market Clearance in the Aggregate

In a general equilibrium context, market clearance requires that the markets for all goods and factors clear simultaneously. The total purchases of any one good by all consumers must equal the total quantity of the good produced, and the total supply of any one factor by all the consumers must equal the total purchases of that factor by all the firms in the economy. Hence,

$$\text{Goods markets: } \sum_{h=1}^{H} X_{hg} = X^g \quad g = 1,\ldots,G \qquad (2.3)$$

$$\text{Factor markets: } \sum_{h=1}^{H} V_{hf} = \sum_{g=1}^{G} r_{gf} \quad f = 1,\ldots,F \qquad (2.4)$$

There are $G + F$ market-clearing equations.

Taken together, Eqns (2.1)–(2.4) provide a complete general equilibrium model of an economy. They comprise all the economic information available to the fictional omniscient social planner who is trying to achieve an efficient allocation of resources and an equitable distribution of income.

EFFICIENCY: THE PARETO-OPTIMAL CONDITIONS

Having specified consumers' preferences, the production technologies, and market clearance, the general equilibrium model is sufficiently detailed to determine the pareto-optimal or efficiency conditions for the economy as a whole. To see how this is done, recall that pareto optimality requires the existence of an allocation of resources such that no one consumer can be made better off by a reallocation of resources without simultaneously making at least one other consumer worse off. The locus of pareto-optimal allocations thus defines a frontier in utility space, the utility-possibilities frontier. Fig. 2.1 illustrates the frontier for the two-person case. The axes are the utility levels achieved by persons 1 and 2, based on one particular utility function for each person that describes their preferences.

A point on the frontier such as A satisfies pareto optimality because an increase in the utility of either person from A requires that the utility of the other person must decrease. Conversely, all points under the frontier, such as point C, cannot be pareto optimal because it is possible to move north, east, or northeast from C. That is, either person can be made better off without the other person being made worse off, or both people can be made better off. The region to the north, east, and northeast of C and bounded by the frontier represents the allocations that are pareto superior to C. Points beyond the frontier, such

3. All production functions are assumed to be continuous, twice differentiable, and well behaved in that their Hessians are negative definite, with all goods and factors infinitely divisible. Notice that our specification of production assumes away intermediate products.

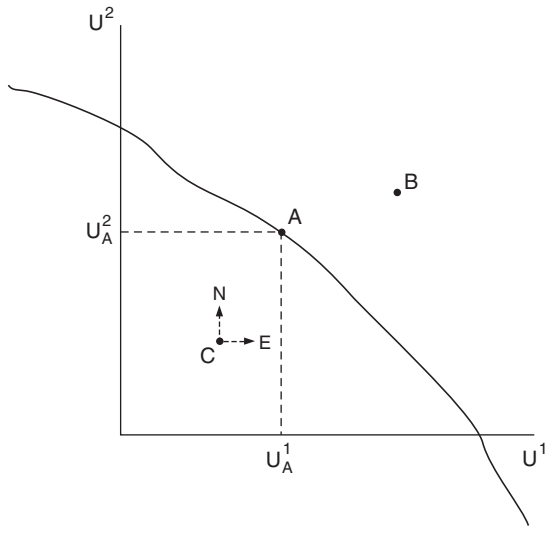

FIGURE 2.1

as B, are simply unattainable, given society's production technologies, individuals' preferences regarding the supply of factors of production, and the requirements of market clearance.

Because the locus of pareto-optimal allocations describes a frontier in utility space, all points on the frontier, such as A, have the following interpretation: Given that person 2 is held at utility level U_A^2, U_A^1 is the maximum satisfaction attainable by person 1. Alternatively, given that person 1 is held at utility level U_A^1, U_A^2 is the maximum utility attainable by person 2.

This interpretation indicates that the set of pareto-optimal allocations for all H individuals can be determined by solving the following problem algebraically: Hold everyone's utility constant except for one person, arbitrarily chosen to be person 1. Maximize person 1's utility subject to the constraints that all other utilities are held constant. Include as additional constraints the G production technologies and the $G+F$ market clearance requirements. Formally,

$$\max_{(X_{hg};V_{hf};X^g;r_{gf})} \quad U^1(X_{1g};V_{1f})$$

$$\begin{aligned} s.t. \quad & \overline{U}^h = U^h(X_{hg};V_{hf}) && h = 2,\ldots,H \\ & X^g = \phi^g(r_{gf}) && g = 1,\ldots,G \\ & \sum_{h=1}^{H} X_{hg} = X^g && g = 1,\ldots,G \\ & \sum_{h=1}^{H} V_{hf} = \sum_{g=1}^{G} r_{gf} && f = 1,\ldots,F \end{aligned}$$

The pareto-optimal conditions follow directly from the first-order conditions of this constrained optimization problem. We will derive them later on in the chapter.

EQUITY: THE SOCIAL WELFARE FUNCTION AND THE OPTIMAL DISTRIBUTION OF INCOME

Although the model as it stands is sufficiently detailed to analyze the necessary conditions for allocational efficiency, it is entirely neutral with respect to any equity norms. Chapter 1 described two types of equity, process equity and end-results equity. The model is silent regarding process equity. This is not so troubling in a social planning context, however, because the planner simply dictates all economic decisions. Process equity norms such as equal opportunity and social mobility are far more relevant in a market context, in which the degree of process equity depends primarily on the structure of the individual markets. Equal opportunity and a reasonable amount of social mobility are likely to be achieved if markets are highly competitive. Market power and other kinds of market imperfections are the chief enemies of these norms.

The same cannot be said about end-results equity, the quest for a just distribution of income. We saw in Chapter 1 that end-results equity is a fundamental issue for any society, even when all the technical and market assumptions for a well-functioning economy hold.

The baseline, social planning efficiency model described above illustrates the end-results equity problem in the following manner. The first-order conditions for the constrained optimum of the model solve for a single allocation of resources, a single point on the utility possibilities frontier. But the constraints imposed upon utility levels of persons $h = 2,\ldots, H$, the $\overline{U}^h$, are entirely arbitrary. Placing at least one of these consumers at a different utility level and solving the model again generates a different allocation of resources, so long as the new constraints permit a feasible solution ($U^1(\,) \geq 0$). Since the utility constraints can be reset in infinitely many ways, solutions to the constrained optimum problem generate an infinity of feasible solutions in general, all points on the utility-possibilities frontier. Furthermore, the model as it stands has no way of choosing a best allocation among these allocations. According to the pareto criterion, all allocations on the frontier are optimal and therefore equivalent. Pareto optimality is an extremely weak normative criterion in this sense.

The inability of the pareto criterion to choose a best allocation is a glaring weakness for a normative theory of the public sector. For instance, the following allocations are equivalent in a two-person economy in terms of the pareto criterion: Person 2 receives almost all the goods and services, and person 1 almost nothing; each person receives an equal allocation of the goods and services; person 1 receives almost all the goods and services, and person 2 almost nothing. The baseline model is completely neutral regarding these outcomes.

Societies are typically not so neutral, however. They embrace a set of end-results equity norms and devise some method of ranking the possible outcomes according to these norms. At the very least, most societies express a concern about the extremes of wealth and poverty.

The Bergson–Samuelson Social Welfare Function

Because most public sector economists believe economic analysis is properly concerned with end-results equity, they have seen fit to include a representation of distributional rankings in their models. The model requires a function that indicates the desirability from society's perspective, the social welfare, of all the possible distributions of individual utility or well-being. The function almost universally chosen for this purpose is the so-called Bergson–Samuelson *individualistic social welfare function*,[4] first described by Abram Bergson and Paul Samuelson in the late 1930s:

$$W = W\left[U^1\left(X_{1g};\ V_{1f}\right), \ldots, U^H\left(X_{Hg};\ V_{Hf}\right)\right]$$

or simply

$$W = W\left[U^h\left(X_{hg};\ V_{hf}\right)\right] \tag{2.5}$$

with $\partial W/\partial U^h > 0$, for all h.

The social welfare function is said to be individualistic because its only arguments are the individuals' utility functions. That is, $W(\ )$ measures the social welfare attained in each possible state of the economy by considering only the utility level or well-being of each individual in that state. Nothing else about the economy matters from a social perspective. Moreover, the individuals themselves determine how well off they are, in keeping with the principle of consumer sovereignty. The Bergson–Samuelson individualistic method of measuring social welfare is therefore consistent with the humanistic view that the goal of an economic system is to promote individual well-being.

The social welfare function gives, in effect, the ethical weight that society confers on each individual in its determination of end-results equity. The ethical weights are usually stated in terms of the first partial derivative of $W(\ )$. $\partial W/\partial U^h$ is the *marginal social welfare weight* for person h, the increase in social welfare resulting from a marginal increase in the utility of person h, holding all other utilities constant.

The condition $\partial W/\partial U^h > 0$, for all h, means that the social welfare rankings honor the pareto principle: If one person's utility increases (decreases), all other utilities held constant, then social welfare must increase (decrease). In other words, all pareto-superior reallocations increase social welfare, and all pareto-inferior reallocations decrease social welfare. Notice, though, that the rankings implied by $W(\ )$ are broader than those implied by the pareto criterion. The function $W(\ )$ can compare two allocations in which a movement from the first to the second increases some utilities while decreasing others' utilities. The pareto criterion cannot make this comparison.

Nobel Laureate Wassily Leontief claims that economists can agree on only two principles of distributive justice, that social welfare should be individualistic and that it should satisfy the pareto principle (Leontief, 1966). His remark underscores the popularity of the Bergson–Samuelson social welfare function among economists, because these are the two properties that they thought a social welfare function should possess.

The Bergson–Samuelson social welfare function completes the baseline model by representing a complete ordering of the well-being of its individual members, analogous to the complete ordering of goods and factors provided by the utility index of an individual consumer. A complete ordering implies that society can make a pairwise, ordinal ranking of all the points in utility space in terms of preference or indifference. It further implies that the ranking is transitive. For example, if point A is preferred to point B, and point B is preferred to point C, then A must be preferred to C. Society cannot solve the problem of end-results equity without a complete ordering of individual outcomes, and the social welfare function is chosen to be consistent with that ordering.

Graphically, $W(\ )$ generates a set of social welfare indifference curves in U^1–U^2 space, depicted by W_0, W_1, and W_2 in Fig. 2.2, having most of the properties associated

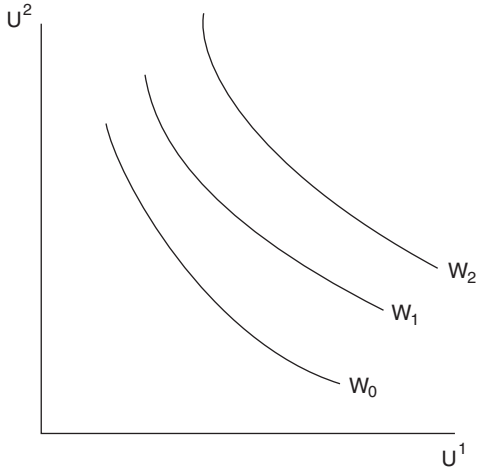

FIGURE 2.2

4. After Abram Bergson and Paul Samuelson, who first described the function. Samuelson used this construct in his 1954 article, "The Pure Theory of Public Expenditure," referred to in footnote 1. Refer to Samuelson's lucid discussion of the social welfare function in Samuelson (1965), pp. 219–230. See also Bergson (1938).

with an individual's indifference curves.[5] The slope of a social welfare indifference curve is the ratio of the marginal social welfare weights of the two individuals.[6]

The objective function of the social planner is to maximize $W(\ )$. In terms of Fig. 2.2, society's goal is to reach the highest possible social welfare indifference curve, just as the consumer's goal is to reach the highest possible indifference curve.

The social welfare function is one of the more convenient analytical constructs in all of economics. It simultaneously solves two of the more difficult normative issues in public sector theory. On the one hand, it represents society's norms regarding end-results equity and thereby answers the distribution question. On the other hand, it resolves the indeterminacy of which of the efficient points society should choose along the utility possibilities frontier.

Refer to Fig. 2.3. The social welfare function selects the distributionally best allocation among the infinity of pareto-optimal allocations along the utility-possibilities frontier. Point *B* represents this distributionally best allocation in the figure, the point at which the utility-possibilities frontier attains the highest numbered social welfare indifference curve.[7] Francis Bator referred to this point as the "bliss point," a name that has stuck in the public sector literature (Bator, 1957). The bliss point maximizes social welfare. As such, it represents a complete solution to the social planners' problem, a solution that best meets the public interest in efficiency and (end-results) equity.

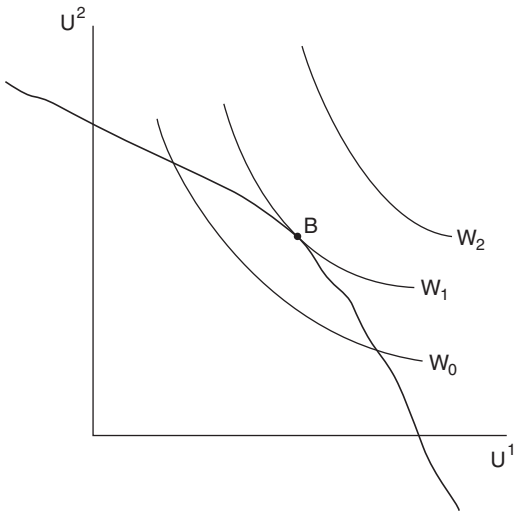

FIGURE 2.3

Limitations of the Social Welfare Function

The analytical usefulness of the social welfare function is clear enough, but its practical significance for policy analysis is very much an open question. Unfortunately, the social welfare function also happens to be one of the more problematic constructs in all of economic theory. We will mention a few of the difficulties here and return to them in more detail in Chapter 3.

The first difficulty is simply trying to determine what the social welfare function is for any nation. The social welfare function is a political concept, not a market concept. It reflects the collective will of the people regarding their notions of distributive justice expressed through the political process. Indeed, the social welfare function is the only explicit element of political content in all of normative public sector economics. The idea of government-as-agent passively representing the desires of the people stops at the social welfare function, because the political process itself is assumed to play a role in shaping the social welfare function.

Deciding what function has evolved from the political process is a difficult question, however. Political signals are often more mixed than market signals and more difficult to test for. Compare, for example, the marginal rate of substitution (MRS) along an individual's indifference curve with the MRS along a social welfare indifference curve. Economists assume that the individual's MRS equals the price ratio of the two goods from the first-order condition for maximizing utility. What, though, is the MRS along a social welfare indifference curve? To what extent is society willing to trade off one person's well-being for another person's well-being on the margin? Has society reached a consensus

5. In particular, the curves are everywhere convex to the origin, society is indifferent among the utility distributions along any one curve, higher numbered curves imply higher levels of social welfare, and no two indifference curves may intersect.

6. The ordinal property of $W(\ )$ deserves comment because the arguments of the social welfare function, unlike those of individual's utility functions, are ordinal. From consumer theory we know that monotonic transformations of an individual's utility function leave the goods demands and factor supplies unchanged. Since these functions themselves are arguments of the social welfare function, arbitrary (monotonic) transformations of the individual's utility functions could easily change the social welfare rankings. But Samuelson and Bergson assumed that if such transformations occurred, the social welfare function would itself change form to preserve the original rankings. There does exist a method of reformulating $W(\ )$ to preserve the individual rankings for any given set of monotonic transformations of the individual utility functions. For a discussion of the transformations that preserve the ordinality of W, see Arrow (1983). Also, Samuelson discusses the ordinal properties of *W* in Samuelson (1981). The interested reader should consult Roemer (1996), for a comprehensive and up-to-date treatment of the social welfare function. Roemer concludes that the arguments of the social welfare function must be something measurable for every individual to make the function fully operational.

7. Since continuity is not required of either $W(U^h)$ or the utility-possibilities frontier, *B* may not be a point of tangency.

on the MRS? If so, how do we test for that MRS? No obvious answers come to mind.

A second difficulty relates to the ethical content of the social welfare function. What should the marginal social welfare weights, $\partial W/\partial U^h$, be for different people? As noted in Chapter 1, no one has come up with a convincing answer to this question. All we have are some suggestions (to be discussed in Chapter 3). This is unsettling, to say the least, since the social welfare function is one of the normative linchpins of economic theory. The marginal social welfare weights are society's norms regarding distributive justice, and a normative theory ought to be able to say something about what those norms should be.

A third difficulty is Arrow's impossibility theorem regarding collective decisions of any kind, also noted in Chapter 1. Arrow's theorem shows that a democratic society may not be able to produce a consistent social welfare function when there is disagreement about the appropriate ethical norms, as there certainly is. A social welfare function may evolve from the political process, but not necessarily in a manner that would be acceptable to a democratic society.

Despite these severe problems, we will follow the conventional practice of using the social welfare function to represent the distributional judgments of society. Societies do care about the distributional implications of their government's policies, and government decision making ought to reflect this concern. Therefore, the prudent course is to incorporate the social welfare function into a general equilibrium model that will be used to develop normative policy rules. This at least allows us to see how the concern for equity might affect the government's decision rules.

At the same time, the social welfare function should not be viewed as anything more than an analytical device representing society's concern for distributive equity. It is not meant to suggest what the distributional judgments should be, other than that they be consistent, individualistic, and satisfy the pareto principle. The alternative of ignoring social welfare rankings entirely because we do not know what they are or should be would simplify the analysis, but it would not produce a meaningful normative theory if society really does care about end-results equity.[8]

MAXIMIZING SOCIAL WELFARE

Adding the social welfare function to the general equilibrium model significantly changes the nature of the model as a foundation for normative policy analysis. The policy objective becomes one of maximizing social welfare, as represented by the social welfare function, rather than simply tracing out the locus of pareto-optimal allocations. Moreover, all individual utilities are allowed to vary, so that the formal model is constrained only by the G production functions and the $G+F$ market clearance equations. The first-order conditions of the model simultaneously determine the set of pareto-optimal and distributional conditions that bring society to the bliss point, the single best allocation and distribution of resources.

Analytically, social welfare maximization is represented as follows:

$$\begin{aligned} \max_{(X_{hg};V_{hf};X^g;r_{gf})} \quad & W\left[U^h\left(X_{hg};V_{hf}\right)\right] \\ s.t. \quad & X^g = \phi^g\left(r_{gf}\right) \qquad g = 1,\ldots,G \\ & \sum_{h=1}^{H} X_{hg} = X^g \qquad g = 1,\ldots,G \\ & \sum_{h=1}^{H} V_{hf} = \sum_{g=1}^{G} r_{gf} \qquad f = 1,\ldots,F \end{aligned}$$

Defining multipliers for each of the constraints and setting up the Lagrangian, the problem becomes

$$\begin{aligned} \max_{(X_{hg};V_{hf};X^g;r_{gf})} L = {} & W\left[U^h\left(X_{hg};V_{hf}\right)\right] \\ & + \sum_{g=1}^{G} \mu_g\left[X^g - \phi^g\left(r_{gf}\right)\right] \\ & + \sum_{g=1}^{G} \delta_g\left[\sum_{h=1}^{H} X_{hg} - X^g\right] \\ & + \sum_{f=1}^{F} \pi_f\left[\sum_{h=1}^{H} V_{hf} - \sum_{g=1}^{G} r_{gf}\right] \end{aligned}$$

Necessary Conditions for Social Welfare Maximization

The first-order conditions for this model are

$$\frac{\partial L}{\partial X_{hg}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hg}} + \delta_g = 0 \quad h = 1,\ldots,H \quad g = 1,\ldots,G \tag{2.6}$$

$$\frac{\partial L}{\partial V_{hf}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial V_{hf}} + \pi_f = 0 \quad h = 1,\ldots,H \quad f = 1,\ldots,F \tag{2.7}$$

$$\frac{\partial L}{\partial X^g} = \mu_g - \delta_g = 0 \quad g = 1,\ldots,H \tag{2.8}$$

8. The comments in this section barely scratch the surface of a voluminous literature on collectively determined decision rules. It is enough for our purposes to establish the central role of the social welfare function in normative public sector analysis. We would recommend Mueller (1976), as a starting point for the student interested in the theory of social choice mechanisms. See also, Arrow et al. (2002).

$$\frac{\partial L}{\partial r_{gf}} = -\mu_g \frac{\partial_\Phi g}{\partial r_{gf}} - \pi_f = 0 \quad g = 1, \ldots, G \quad f = 1, \ldots, F \tag{2.9}$$

and the constraints are

$$X^g = \phi^g\left(r_{gf}\right) \quad g = 1, \ldots, G \tag{2.10}$$

$$\sum_{h=1}^{H} X_{hg} = X^g \quad g = 1, \ldots, G \tag{2.11}$$

$$\sum_{h=1}^{H} V_{hf} = \sum_{g=1}^{G} r_{gf} \quad f = 1, \ldots, F \tag{2.12}$$

There are $HG + HF + GF + 3G + F$ equations in all, which we assume generate a unique solution to the $HG + HF + GF + 3G + F$ variables of the model, consisting of the $HG + HF + GF + G$ economic variables, X_{hg}, V_{hf}, r_{gf}, X^g, and the $2G + F$ Lagrangian multipliers.[9]

The First-Best Efficiency-Equity Dichotomy

A most useful feature of these equations for policy purposes is that the first $(HG + HF + G + GF)$ first-order conditions can be combined into two distinct sets. One set contains the pareto-optimal conditions, the necessary conditions for an efficient allocation of resources. The other set contains the interpersonal equity conditions, the necessary conditions for an optimal distribution. The pareto-optimal conditions do not contain any social welfare terms, whereas the interpersonal equity conditions do. This makes intuitive sense considering that the pareto-optimal conditions describe how to achieve the allocations that bring the economy to the utility-possibilities frontier, and we know that they can be determined using a model that does not employ a social welfare function. The interpersonal equity conditions, in contrast, must involve the social welfare function, since that function contains the additional ethical information needed to determine the optimal distribution.

The pareto-optimal conditions themselves divide into three distinct sets: one describing the optimal consumption conditions, one describing the optimal production conditions, and one describing the optimal interrelationships between production and consumption.

To obtain the optimal consumption conditions, standardize on any one person and consider the following pairs of first-order conditions:

1. Any two goods demanded by that person.
2. Any two factors supplied by that person.
3. Any one good demanded and any one factor supplied by that person.

Pairing the first-order conditions in this manner eliminates any terms involving the social welfare function.

Since production does not involve the social welfare function, all pairs of production relationships generate pareto-optimal conditions, including:

4. Any one factor used in the production of any two goods.
5. Any two factors used in the production of any one good.

The interrelationships between production and consumption are derived by combining the first two sets of pairings. There are three relevant combinations:

6. The rate at which any one person is willing to trade any two goods (P1) with their efficient rate of exchange in production (P4).
7. The rate at which any one consumer is willing to substitute any two factors (P2) with their efficient rate of exchange in production (P5).
8. The rate at which any one consumer is willing to substitute any one good for any one factor (P3) with their efficient rate of exchange in production (P4).

Taken together, these eight pairings generate all the conditions necessary for the economy to be on its utility-possibilities frontier. Should any one of them fail to hold, the omniscient planner can always find a reallocation of resources that will increase the utility of at least one person without making any other person worse off.

To derive the interpersonal equity conditions, the first-order conditions must be paired in such a way as to retain the social welfare terms. Since these terms involve the consumers, there are only two possible ways of doing this. Compare:

1. Any one good demanded by two different people.
2. Any one factor supplied by two different people.

A final point worth noting by way of an introduction to policy analysis is that this dichotomization of the first-order conditions is not peculiar to the baseline general equilibrium model. As we shall see, it applies to all general equilibrium social planning models that assume government policy is not constrained in any way other than by the fundamental elements of any economy: preferences, production technologies, and market clearance. Policy analysis under this assumption is called *first-best analysis.* This feature is extremely important as a practical matter because it implies that the government can pursue its equity and

9. Existence of a unique solution is never guaranteed by simply matching the number of equations with the number of variables, but we do not want to consider the problem of existence in the text. Hence, existence of a unique solution for all maximization problems will be assumed throughout.

efficiency goals with distinct sets of policy tools. We will return to this point in Chapter 3.

The Pareto-Optimal Conditions

To demonstrate the derivation and interpretation of the pareto-optimal conditions, we will consider the three conditions most commonly presented in microeconomic analysis, corresponding to the pairings in 1, 5, and 6 above. If all factors of production are supplied by consumers in absolutely fixed amounts, then these conditions are the only necessary conditions for a pareto optimum. The pairings 2, 3, 7, and 8 have no meaning when factor supplies are fixed because the fixed factors are not decision variables for the consumers. In general, however, all eight conditions are necessary for overall economic efficiency.

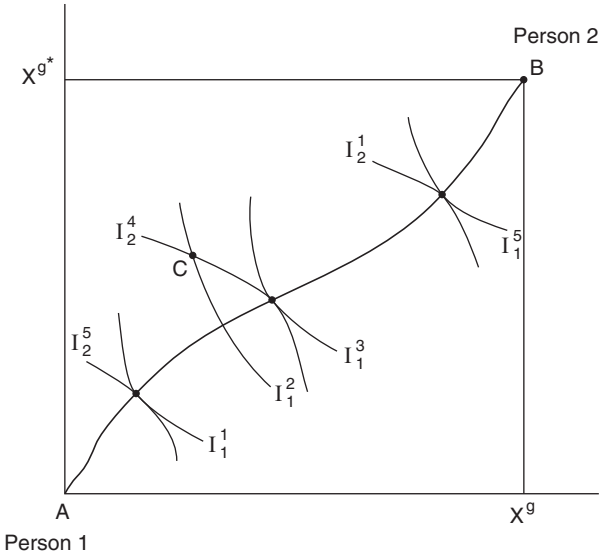

FIGURE 2.4

Condition P1 (= Marginal Rate of Substitution)

Consider the first-order conditions for any two goods demanded by any one person, say X_{hg} and X_{hg^*}:

$$\frac{\partial L}{\partial X_{hg}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hg}} + \delta_g = 0 \tag{2.13}$$

$$\frac{2L}{\partial X_{hg^*}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hg^*}} + \delta_{g^*} = 0 \tag{2.14}$$

Dividing Eqn (2.13) by (2.14) yields

$$\frac{\dfrac{\partial U^h}{\partial X_{hg}}}{\dfrac{\partial U^h}{\partial X_{hg^*}}} = \frac{\delta_g}{\delta_{g^*}} \qquad \begin{matrix}\text{all } h = 1,\ldots,H \\ \text{any } g, g* = 1,\ldots,G\end{matrix} \tag{2.15}$$

Notice that the social welfare term $\partial W/\partial U^h$ cancels on the left-hand side (LHS) of Eqn (2.15), so that the LHS is the familiar MRS between goods g and g^* for person h. Also, the right-hand side (RHS) of Eqn (2.15) is independent of h. Therefore, condition P1, Eqn (2.15), says that the MRS between any two goods must be the same for all people.

To represent this condition geometrically, consider an economy with two people, persons 1 and 2, and two goods, X^g and X^{g^*}. Fig. 2.4 is the Edgeworth box for which the axes are society's total production of X^g and X^{g^*}. Person 1's indifference curves are drawn with reference to the lower left-hand corner as the origin, and person 2's indifference curves are drawn with reference to the upper right-hand corner as the origin. The equality of marginal rates of substitution is represented by the contract curve AB, the locus of points at which the two sets of indifference curves are tangent. Any point along the contract curve is efficient. Any allocation off the contract locus, such as C, is inefficient since some other allocation exists that can make one or both people better off without making anyone else worse off. For example, suppose at C the slopes of the indifference curves are such that $\text{MRS}^1_{X^g,X^{g^*}} = 2$ and $\text{MRS}^2_{X^g,X^{g^*}} = 1.8$. If the social planner forces person 1 to give 1.9 units of X^{g^*} to person 2 in exchange for 1 unit of X^g, person 1 is better off, since he or she is willing to exchange at a 2-for-1 ratio, by the definition of the MRS. Person 2 will accept the 1.9-for-1 exchange as well, since he or she is willing to trade 1 unit of X^g for only 1.8 units of X^{g^*} in return. Any (small) trade between the ratios 2:1 and 1.8:1, including the boundaries, generates an allocation of the goods that is pareto superior to C (at the trade boundaries, only one person gains, but the other is no worse off).

Only when the two MRS are equal is no such beneficial trade possible, which is true for any point along the contract curve. Note, finally, that the pareto criterion cannot rank points along the contract curve—they are all pareto optimal by condition P1.

Condition P5 (= Marginal Rate of Technical Substitution)

Consider any two factors used in the production of any one good, say r_{gf} and r^*_{gf}:

$$\frac{\partial L}{\partial r_{gf}} = -\mu_g\frac{\partial \phi^g}{\partial r_{gf}} - \pi_f = 0 \tag{2.16}$$

$$\frac{\partial L}{\partial r_{gf^*}} = -\mu_g\frac{\partial \phi^g}{\partial r_{gf^*}} - \pi_{f^*} = 0 \tag{2.17}$$

Dividing Eqn (2.16) by (2.17) yields

$$\frac{\dfrac{\partial \phi^g}{\partial r_{gf}}}{\dfrac{\partial \phi^g}{\partial r_{gf^*}}} = \frac{\pi_f}{\pi_{f^*}} \qquad \begin{array}{l}\text{all } g = 1, \ldots, G \\ \text{any } f, f^* = 1, \ldots, F\end{array} \tag{2.18}$$

The LHS of Eqn (2.18) is the marginal rate of technical substitution (MRTS) of factors f and f^* in the production of good g.[10] The RHS of Eqn (2.18) is independent of g. Therefore, condition P5, Eqn (2.18), states that the MRTS between any two factors in the production of a good must be equal for all goods. The usual way of representing this condition geometrically is to think of the factors f and f^* as capital (K) and labor (L) and draw a production box analogous to the Edgeworth consumption box, as in Fig. 2.5.

The axes represent society's total supply of capital and labor, a representation possible only under the assumption of fixed factor supplies. The isoquants $q_g^1, \ldots, q_g^5$ for X^g are drawn with reference to the lower left-hand corner as the origin, and the isoquants $q_{g^*}^1, \ldots, q_{g^*}^5$ for X^{g^*} are drawn with reference to the upper right-hand corner. As before, the contract locus of tangency points represents the pareto-optimal allocations of K and L between the two goods, X^g and X^{g^*}, and all points off this locus are dominated according to the pareto criterion by some point on the locus. The pareto criterion is defined in terms of production in this context, but production efficiency is necessary for full pareto optimality defined in terms of individuals' utilities. If society can produce more of at least one good without sacrificing production of some other good, then the planner can distribute the bonus to make someone better off without making anyone else worse off.

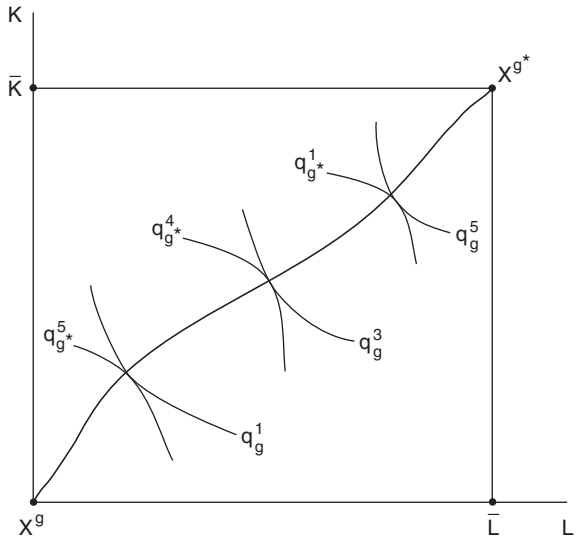

FIGURE 2.5

The contract locus in factor space in turn bears a point-to-point correspondence with the production-possibilities frontier in goods space, depicted in Fig. 2.6. If society is producing along the contract locus in factor space, it cannot realign its resources to produce more of one good without sacrificing some of the other good. But, this is exactly what the production-possibilities frontier represents, the locus of pareto-efficient production of the goods.

Condition P6 (MRS = Marginal Rate of Transformation)

Pareto optimality requires that the rate at which consumers are willing to trade any one good for any other equal their rate of transformation in (efficient) production. The slope of the production-possibilities frontier in Fig. 2.6 is the marginal rate of transformation (MRT) between the two goods, X^g and X^{g^*}, in production, assuming efficient production. To derive the MRT algebraically, consider a single factor f switched from the production of good X^g to good X^{g^*}.

The first-order conditions for r_{gf} and r_{g^*f} are

$$\frac{\partial L}{\partial r_{gf}} = -\mu_g \frac{\partial \phi^g}{\partial r_{gf}} - \pi_f = 0 \tag{2.19}$$

$$\frac{\partial L}{\partial r_{g^*f}} = -\mu_{g^*} \frac{\partial \phi^{g^*}}{\partial r_{g^*f}} - \pi_f = 0 \tag{2.20}$$

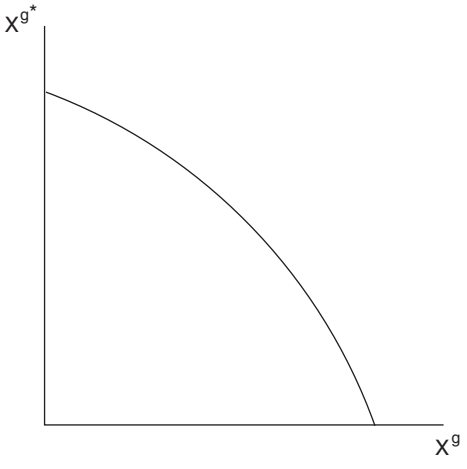

FIGURE 2.6

10. Notice that the numerator and denominator of Eqn (2.18) equal the marginal products of factors f and f^* in the production of g. Hence, the marginal rate of technical substitution between any two factors is the ratio of their marginal products.

Therefore,

$$-\mu_g \frac{\partial \phi^g}{\partial r_{gf}} = -\mu_{g^*} \frac{\partial \phi^{g^*}}{\partial r_{g^* f}} \tag{2.21}$$

or

$$\frac{\dfrac{\partial \phi^{g^*}}{\partial r_{g^* f}}}{\dfrac{\partial \phi^g}{\partial r_{gf}}} = \frac{\mu_g}{\mu_{g^*}} \quad \begin{matrix} \text{all } f = 1, \ldots, F \\ \text{any } g, g^* = 1, \ldots, G \end{matrix} \tag{2.22}$$

The LHS of Eqn (2.22) is the MRT between X^g and X^{g^*} obtained by switching factor f from good X^g to good X^{g^*}. Since the RHS of Eqn (2.22) is independent off, Eqn (2.22) holds for all factors switched between X^{g^*} and X^g. Thus, the LHS is simply the MRT between X^{g^*} and X^g. (Eqn (2.22) is also production condition P4.)

The $\text{MRT}_{g^*,g}$ must now be related to each consumer's $\text{MRS}_{g^*,g}$. From the consumption condition P1, Eqn (2.15),

$$\frac{\dfrac{\partial U^h}{\partial X_{hg}}}{\dfrac{\partial U^h}{\partial X_{hg^*}}} = \frac{\delta_g}{\delta_{g^*}} = \text{MRS}^h_{g^*,g} \quad \begin{matrix} \text{all } h = 1, \ldots, H \\ \text{any } g^*, g = 1, \ldots, G \end{matrix} \tag{2.23}$$

Consider, next, the first-order conditions with respect to X^g, the aggregate production of good g:

$$\frac{\partial L}{\partial X^g} = \mu_g - \delta_g = 0 \quad g = 1, \ldots, G \tag{2.24}$$

Thus, $\mu L_g = \delta_g$, $g = 1,\ldots, G$, so that

$$\frac{\dfrac{\partial U^h}{\partial X_{hg}}}{\dfrac{\partial U^h}{\partial X_{hg^*}}} = \frac{\dfrac{\partial \phi^{g^*}}{\partial r_{g^* f}}}{\dfrac{\partial \phi^g}{\partial r_{gf}}} \quad \text{any } g^*, g = 1, \ldots, G \tag{2.25}$$

In other words,

$$\text{MRS}^h_{g*,g} = \text{MRT}_{g*,g} \quad \text{any } g*, g = 1, \ldots, G \tag{2.26}$$

To picture this result, suppose society is at point A on the production-possibilities frontier in Fig. 2.7. Let point A define the dimensions of an Edgeworth consumption box placed inside the frontier, consisting of $X^{g^*}_A$ units of X^{g^*} and X^g_A units of X^g.

Condition P6, Eqn (2.25), says that society must distribute the total product at A between persons 1 and 2 such that the common MRS between the two goods equals their MRT in production. Of all the pareto-efficient points on the consumption contract curve, society must choose A′, giving person 1 $(X^{A'}_{1g}, X^{A'}_{1g^*})$ and person 2 the remainder.[11]

11. There may be no point that satisfies Eqn (2.26) given A or many points.

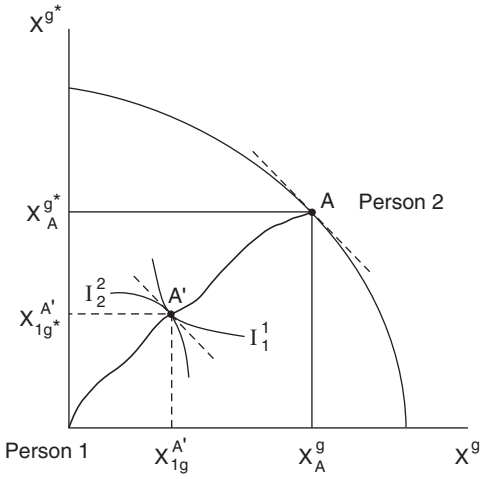

FIGURE 2.7

Notice that while condition P6 has distributional implications, it is not a distributional rule in the sense of an interpersonal equity condition because it does not involve the social welfare function. The distribution $[(X^{A'}_{1g},\ X^{A'}_{1g^*}),\ (X^g_{A'} - X^{A'}_{1g}, X^{g^*}_{A'} - X^{A'}_{1g^*})]$ is not determined by interpersonal utility comparisons.

Having satisfied P1, P5, and P6 simultaneously, A' defines a single point on the utility-possibilities frontier, point A'' in Fig. 2.8, corresponding to A' in Fig. 2.7. ($U^2_{A''}$ in Fig. 2.8 is the utility achieved by person 2 on indifference curve I^2_2 in Fig. 2.7. $U^1_{A''}$ in Fig. 2.8 is the utility achieved by person 1 on indifference curve I^1_1 in Fig. 2.7.) Thus, conditions P1, P5, and P6 are consistent with an infinity of allocations.

If factor supplies are variable, attaining the utility-possibilities frontier requires satisfying four additional pareto-optimal conditions, corresponding to the pairings

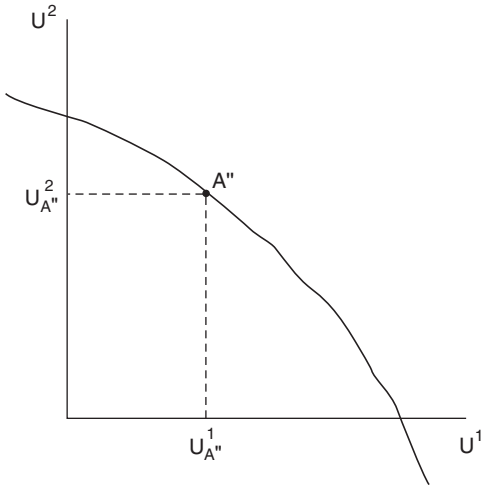

FIGURE 2.8

of first-order conditions 2, 3, 7, and 8. They are derived following the same procedures used to generate conditions P1, P4, P5, and P6, an exercise that will be left to the reader.

The conditions are as follows:

P2: The MRS between any two factors in supply must be equal for all people.
P3: The MRS between a good and a factor must be equal for all consumers.
P7: The common MRS between any two factors in supply must equal their common MRTS in the production of any good.
P8: The common MRS between any good demanded and any factor supplied must equal the marginal product of that factor in producing that good (or the MRTS between the good and the factor in production).

Pareto Optimality and Perfect Competition

The first fundamental theorem of welfare economics states that if all the technical assumptions listed in Chapter 1 hold, then a perfectly competitive market system generates all eight necessary conditions for full pareto optimality. A formal proof of the theorem requires mathematical techniques beyond the scope of this text, but an intuitive, heuristic argument illustrating the theorem is relatively straightforward. As with the derivation of the conditions themselves, we will illustrate this theorem with reference only to conditions P1, P5, and P6.

That condition P1 is satisfied in a competitive market economy follows immediately from the behavioral assumption that consumers maximize utility subject to their budget constraints and the fact that in a perfectly competitive economy all consumers are price takers facing the same set of prices. Under these conditions, each utility-maximizing consumer sets the MRS between any two goods equal to the ratio of their prices.[12] If all consumers do this, and each faces the same set of prices, then the MRS between any two goods must be equal for all consumers.

Similarly, condition P5 follows directly from the fact that profit-maximizing firms produce any given output with the least cost combination of factors of production. If a firm cannot influence factor prices, then it minimizes cost by producing such that the MRTS between any two of its factors equals the ratio of the factor prices.[13] If markets are perfectly competitive, then all firms will face the same set of factor prices. Consequently, the MRTS between any two factors is equalized throughout the economy, as required by condition P5.

Condition P6 follows from the result that, in competitive markets, firms produce the output at which price equals marginal cost to maximize profit. If $p_g = \mathrm{MC}_g$ and $p_{g^*} = \mathrm{MC}_{g^*}$, then,

$$\frac{p_g}{p_{g^*}} = \frac{\mathrm{MC}_g}{\mathrm{MC}_{g^*}} \quad \text{any } g, g^* = 1, \ldots, G \tag{2.27}$$

Each consumer (h) sets $\mathrm{MRS}^h_{g^*g} = p_g/p_{g^*}$. Moreover, assuming efficient production (that conditions P4 and P5 hold), the ratio of marginal costs between any two goods is equal to their MRT. MC_g gives the extra cost of (efficiently) producing an extra unit of X^g, and similarly for MC_{g^*}. Hence, the ratio $\mathrm{MC}_g/\mathrm{MC}_{g^*}$ gives the rate at which g^* substitutes for g in production by transferring a dollar's worth of resources from g to g^*, or vice versa.[14] Therefore, with marginal cost pricing in every market, $\mathrm{MRT}_{g^*,g} = p_g/p_{g^*}$, and condition P6 is satisfied for all goods and services.

That perfectly competitive markets also generate conditions P2, P3, P7, and P8 when factor supplies are variable can be shown by similar reasoning.

12. Formally, each consumer h solves the following problem:

$$\underset{(X_{hg}, V_{hf})}{\mathrm{Max}} \; U^h(X_{hg}, V_{hf})$$
$$s.t. \sum_{g=1}^{G} p_g X_{hg} + \sum_{f=1}^{F} w_f V_{hf} = 0$$

where

p_g = the price of the gth good.
w_f = the price of the fth factor.

The first-order conditions for any two goods g and g^* imply

$$\frac{\frac{\partial U^h}{\partial X_{hg}}}{\frac{\partial U^h}{\partial X_{hg^*}}} \equiv \mathrm{MRS}^h_{g^*,g} = \frac{p_g}{p_{g^*}} \quad \text{all } g, g^* = 1, \ldots, G$$

13. Formally, each firm (g) solves the following problem:

$$\min_{(r_{gf})} \sum_{f=1}^{F} w_f r_{gf}$$
$$s.t. \; X^g = \phi(r_{gf})$$

The first-order conditions for any two factors f and f^* imply

$$\frac{\frac{\partial \phi^g}{\partial r_{gf}}}{\frac{\partial \phi^g}{\partial r_{gf^*}}} \equiv \mathrm{MRTS}^g_{f^*,f} = \frac{w_f}{w_{f^*}} \quad \text{all } f^*, f = 1, \ldots, F$$

14. That the marginal rate of transformation between g and g^* is equal to the ratio of their marginal costs follows immediately from Eqn (2.22). Switch a dollar of factor f from g^* to g. The numerator and denominator measure the per dollar loss and gain in outputs g^* and g, respectively. Inverting each term gives the ratio of marginal costs.

The Interpersonal Equity Conditions

The competitive market system can generate the full set of pareto-optimal conditions, but no more. Like the pareto criterion itself, the market is neutral regarding the points on the utility-possibilities frontier. If society is not neutral, clearly preferring some distributions of the economy's goods and services to others, it must ask the government to carry out its collective will with respect to the distribution. Assuming that the Bergson–Samuelson social welfare function represents its distributional norms and society wants to maximize social welfare, the government must act according to the dictates of two additional sets of first-order conditions, the interpersonal equity conditions. The interpersonal equity conditions combine with the pareto-optimal conditions to bring the economy to the bliss point on the utility-possibilities frontier.

As indicated above, the interpersonal equity conditions arise from pairings of the first-order conditions Eqns (2.6) and (2.7) that standardize on a single good or factor. Consider condition IE1, a single good demanded by two different people (say, X_{hg} and X_{h^*g}). The first-order conditions are

$$\frac{\partial L}{\partial X_{hg}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hg}} + \delta_g = 0 \tag{2.28}$$

$$\frac{\partial L}{\partial X_{h^*g}} = \frac{\partial W}{\partial U^{h^*}}\frac{\partial U^{h^*}}{\partial X_{h^*g}} + \delta_g = 0 \tag{2.29}$$

Therefore,

$$\begin{aligned}\frac{\partial W}{\partial U^1}\frac{\partial U^1}{\partial X_{1g}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hg}} = \cdots = \frac{\partial W}{\partial U^{h^*}}\frac{\partial U^{h^*}}{\partial X_{h^*g}} = \cdots \\ = \frac{\partial W}{\partial U^H}\frac{\partial U^H}{\partial X_{Hg}} = -\delta_g \; g = 1, \ldots, G\end{aligned} \tag{2.30}$$

$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hg}}$ is the *social marginal utility of consumption* of good g for person h, equal to the product of the marginal social welfare weight of person h, $\partial W/\partial U^h$, and the private marginal utility of consumption of good g of person h, $\partial U^h/\partial X_{hg}$. It indicates the marginal increase (decrease) in social welfare from a *ceteris paribus* unit increase (decrease) in person h's consumption of good g. Condition (2.30) says that interpersonal equity is achieved only if all goods are distributed such that, on the margin, the increase in social welfare is the same no matter who consumes the last unit of the good. A similar condition applies to all factor supplies as well.[15] By following this decision rule and assuming the pareto-optimal conditions are satisfied, society in effect moves along the utility-possibilities frontier to the bliss point, which is distributionally the best of all possible pareto-optimal allocations.

Three policy implications of this rule should be noted.

How Many Goods to Redistribute?

First, there are not really $(G + F)$ *independent* conditions, one for each good and factor. To the contrary, if the pareto-optimal conditions hold and society is able to satisfy the interpersonal equity condition for any one good g, then the interpersonal equity condition is automatically satisfied for all other goods and factors. To see this, suppose that interpersonal equity holds for good g, so that

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hg}} = \frac{\partial W}{\partial U^{h^*}}\frac{\partial U^{h^*}}{\partial X_{h^*g}} \quad \text{any } h, h^* = 1, \ldots, H \tag{2.31}$$

Assume, also, that pareto-optimal condition P1 holds for goods g and g^*:

$$\frac{\dfrac{\partial U^h}{\partial X_{hg}}}{\dfrac{\partial U^h}{\partial X_{hg^*}}} = \frac{\dfrac{\partial U^{h^*}}{\partial X_{h^*g}}}{\dfrac{\partial U^{h^*}}{\partial X_{h^*g^*}}} \quad \begin{matrix}\text{any } h, h^* = 1, \ldots, H \\ \text{any } g, g^* = 1, \ldots, G\end{matrix} \tag{2.32}$$

or

$$\text{MRS}^h_{g^*,g} = \text{MRS}^{h^*}_{g^*,g} \quad \text{any } h, h^* = 1, \ldots, H \tag{2.33}$$

Restore the social welfare terms in condition P1 (from Eqns (2.13) and (2.14)), maintaining the equality:

$$\frac{\dfrac{\partial W}{\partial U^h}\dfrac{\partial U^h}{\partial X_{hg}}}{\dfrac{\partial W}{\partial U^h}\dfrac{\partial U^h}{\partial X_{hg^*}}} = \frac{\dfrac{\partial W}{\partial U^{h^*}}\dfrac{\partial U^{h^*}}{\partial X_{h^*g}}}{\dfrac{\partial W}{\partial U^{h^*}}\dfrac{\partial U^{h^*}}{\partial X_{h^*g^*}}} \quad \begin{matrix}\text{any } h, h^* = 1, \ldots, H \\ \text{any } g, g^* = 1, \ldots, G\end{matrix} \tag{2.34}$$

15. From conditions (2.7),

$$\frac{\partial L}{\partial V_{hf}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial V_{hf}} + \pi_f = 0$$

$$\frac{\partial L}{\partial V_{h^*f}} = \frac{\partial W}{\partial U^{h^*}}\frac{\partial U^{h_*}}{\partial V_{h^*f}} + \pi_f = 0$$

Hence,

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial V_{hf}} = -\pi_f \quad \begin{matrix}\text{all } h = 1, \ldots, H \\ \text{any } f = 1, \ldots, F\end{matrix}$$

The numerators of the two ratios are equal from the interpersonal equity condition for good g. Therefore, the denominators are also equal, and interpersonal equity is satisfied for g^* as well. Since the choice of g^* was entirely arbitrary, interpersonal equity must hold for all goods if it holds for the gth good.[16]

Thus, the government's task is much easier than it first appears to be. Difficult as it may be to satisfy any of the interpersonal equity conditions, at least they need be satisfied for only one good (or, alternatively, only one factor) in an otherwise competitive economy.

Lump-Sum Redistributions

The second policy implication relates to the actual policies required to satisfy these conditions. The competitive market system is of no help. The interpersonal equity conditions will not hold in general at a competitive general equilibrium, and, if they do not, no natural market forces are at work to bring about the necessary equality. The government must find some other means of satisfying the interpersonal equity conditions. By the same token, government redistributions must not undermine the considerable achievement of the competitive market system, namely, the attainment of full pareto optimality. If social welfare is to be maximized, the government must use the information contained in the social welfare function to move society along the utility-possibilities frontier to the bliss point. It cannot take society inside the frontier.

Only one form of redistribution ensures that the pareto-optimal conditions continue to hold. The redistributions must be *lump sum*, meaning that the amount of the good or factor redistributed among the consumers is invariant to the economic decisions of all consumers and producers. An example is a tax or transfer based on a person's age. The tax liabilities under an age tax are clearly invariant to any economic decisions the taxpayers might make.

Another way to define a lump-sum tax or transfer is to say that it does not distort the operation of the market economy. A tax (transfer) is nondistorting if it does not introduce any inefficiency into the economy, that is, it does not drive the economy beneath its utility-possibilities frontier. For this to be true, the tax (transfer) must allow all the pareto-optimal conditions to hold. But this in turn requires that all consumers and producers face the same prices for the same goods or factors; otherwise, some of the pareto-optimal conditions will not hold. Conversely, taxes (transfers) distort economic decisions by causing different agents to face different prices for the same goods or factors.

An age tax is nondistorting by this definition. Two consumers may pay different amounts of tax under an age tax, but they continue to face the same price ratios for all goods and factors. Therefore, their marginal rates of substitution remain equal for all goods and factors, as required for pareto optimality. In contrast, suppose the government redistributes income using a set of taxes and transfers based on wage income, and consider the tax on wages. The tax drives a wedge between the price of labor paid by the firms and the price of labor received by the consumers. Firms look at the wage *including the tax* when deciding how many workers to hire, whereas workers look at the wage *net of the tax* (their take home pay) when deciding how much labor to supply. Consequently, pareto-optimal conditions P7 and P8 cannot be fully satisfied in the market exchange of labor.

Notice two qualities that lump-sum taxes and transfers *do not* possess. First, it is not true that lump-sum redistributions have no effect on economic activity. Any redistribution program has income effects that tend to change individuals' demands for goods or supplies of factors, with obvious repercussions throughout the entire economy. Second, it is not true that lump-sum redistributions have no effect on the values of the consumers' marginal rates of substitution, producers' marginal rates of technical substitution, and the marginal rates of transformation in production. Prices change in general as demand (and factor supply) curves shift. Therefore, the values of some of the marginal rates of exchange change as well, as consumers and producers equate these margins to relative prices. For instance, the movement along the utility-possibilities frontier occasioned by the government's lump-sum redistribution policy also moves society along its production-possibilities frontier. Since the MRT is the slope of this frontier, marginal rates of transformation necessarily change if the frontier is anything but constant cost (a straight line). Subsequently, all marginal rates of substitution have to change as competitive market forces reestablish the equality between consumers' marginal rates of substitution and the marginal rates of transformation. Lump-sum redistributions only ensure that the pareto-optimal conditions continue to hold, not that they hold at any particular value.[17] A lump-sum redistribution of one of the goods or factors, then, is the absolute minimum policy

16. The F additional interpersonal equity conditions for the variable factor supplies will also be satisfied. This follows immediately from the subset of pareto-optimal conditions in P3 relating the marginal rate of substitution between good g and any factor f and the interpersonal equity condition for good g.

17. A potential confusion on this point arises from the typical exercises in consumer theory that represent lump-sum taxes and transfers as parallel shifts in the consumer's budget line. The parallel shift does not change the consumer's MRS in the new equilibrium. This representation is valid in a general equilibrium context only if the tax or transfer is so small that it has no effect on the overall economy, for example, if that consumer is only one being taxed or receiving a transfer. Any large tax-transfer redistribution changes prices throughout the economy and causes all consumers' budget lines to rotate. There is no distortion from these price changes, however, since consumers and producers face the same new price ratios.

required of the government even in a world of perfect markets with all the technical assumptions of Chapter 1 holding, so long as society cares about end-results equity.

The Social Marginal Utility of Income

One final point about the interpersonal equity conditions deserves mention. Economists typically refer to the interpersonal equity conditions in terms of "income." The relevant social marginal utilities are written as $\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial Y^h}$, where Y^h is the income of person h, and are referred to as the *social marginal utility of income* of person h. The social marginal utility of income is a product of the marginal social welfare weight ($\partial W/\partial U^h$) and the private marginal utility of income ($\partial U^h/\partial Y^h$). The single required interpersonal equity condition is then stated as equalizing the social marginal utilities of income across all individuals and is achieved with lump-sum redistributions of income.

This interpretation of the interpersonal equity condition can be confusing, however, because the meaning of "income" is ambiguous if more than one variable factor is being supplied by consumers. Furthermore, the interpersonal equity conditions of social welfare maximization seem to suggest that physical quantities of some good or factor must be transferred rather than a dollar value of "income." What, then, is the "income" that is being redistributed lump sum?

One possible interpretation is to assume that all consumers possess an initial endowment of some good, say, X^g, which is also produced and sold by some of the firms in the economy. Some consumers may want to consume their entire endowment of X^g and purchase additional quantities either from other consumers or the producers of X^g. Other consumers may consume only a part of their endowment and sell the rest. If the government redistributes the initial endowments, the redistribution is clearly lump sum. If it continues to redistribute until

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hg}} = -\delta_g, \quad \text{for all } h = 1, \ldots, H \tag{2.35}$$

then the interpersonal equity condition for X^g is satisfied. Assuming a competitive market system with all technical assumptions holding, full pareto optimality is also maintained. Hence, the interpersonal equity conditions are satisfied for all other goods and factors as well. Finally, by evaluating the endowments at either the pre- or posttransfer prices of X^g one can speak of transferring a dollar value of "income," or purchasing power.[18]

Another common interpretation is to associate the "income" with some factor of production that consumers supply in absolutely fixed amounts, such as their land holdings. Transferring physical or dollar amounts of this resource is obviously lump sum, since by definition it is not a decision variable of any consumer. Moreover, these transfers move society along its utility-possibilities frontier as those taxed lose utility and those receiving transfers gain utility. In effect, the government is satisfying the interpersonal equity conditions indirectly. Presumably there exists a redistribution of the fixed resource that satisfies the interpersonal equity condition for one of the variable goods or factors (say, X^g). But if $(\partial W/\partial U^h)(\partial U^h/\partial X_{hg}) = -\partial_g$, all $h = 1,\ldots, H$, and pareto optimality holds, then the interpersonal equity conditions hold for all variable goods and factors. Thus, the existence of a fixed factor gives the government the leverage it needs to satisfy the interpersonal equity conditions, even though they are defined in terms of the variable goods and factors.

Finally, it may simply be assumed that the good or factor being transferred is serving as the numeraire, such that its price is equal to one at any general equilibrium. Competitive market economies determine pareto-optimal allocations of resources in terms of relative prices; the absolute price level is entirely arbitrary. Thus, it is always possible to single out a good or factor, set its price equal to one, and solve for the values of all other prices in terms of the one fixed price. If the numeraire good is chosen for redistribution, unit transfers of it are equivalent to unit transfers of purchasing power or "real" income. This is the most general interpretation of "income" and the most common one.

One final comment on equity is in order, a reminder pertaining to the goal of process equity. The interpersonal equity conditions have nothing to do with process equity norms; they relate strictly to the goal of end-results equity, of achieving a just distribution of income. As noted earlier, the competitive market system is relied on to achieve process equity by promoting equal opportunity and social mobility. Our baseline, social planner model has nothing explicit to say regarding process equity, as is true of most models used in public sector economics.

POLICY IMPLICATIONS AND CONCLUSIONS

The principal task in Chapter 2 was to present a baseline version of the standard general equilibrium model used in normative public sector analysis. Nonetheless, the discussion of the interpersonal equity conditions and lump-sum redistributions generated a number of fundamental propositions relating to the goal of end-results equity:

1. If society cares about distributive equity, it must establish a government to carry out its wishes. A perfectly functioning competitive market economy generates an

18. The same analysis could be applied to the endowment of a primary factor, such as inherited capital.

efficient (pareto-optimal) allocation of resources, but even the most perfect market system is neutral regarding the question of end-results equity.

2. Society's norms regarding distributive justice can be represented analytically by a Bergson–Samuelson individualistic social welfare function, whose arguments are the utility functions of each individual in the society. The partial derivative, $\partial W/\partial U^h$, is the marginal social welfare weight, society's ethical judgment about the effect on social welfare of a marginal change in the well-being of person h. The social welfare function comes from the political process. As such, it is the only explicit political content in normative public sector theory.
3. In the best of all worlds, with all the technical assumptions of a well-functioning market system holding and perfectly competitive markets, distributive equity is achieved by a set of lump-sum redistributions satisfying the first-order interpersonal equity conditions of social welfare maximization. The interpersonal equity conditions require that the social marginal utilities of any one good or factor be equalized across all individuals.

The interpersonal equity conditions represent a complete normative theory of the optimal income distribution and redistribution in this setting. The normative question—What is the optimal distribution of resources?—has a remarkably simple answer, in principle. It is the distribution that satisfies the interpersonal equity conditions, given the distributional rankings implied by the underlying social welfare function. If some other distribution happens to exist, then the interpersonal equity conditions provide a complete normative policy prescription for redistributing resources lump sum to achieve the optimal distribution. Nothing more need be said about the government's redistributive policies.

The second theorem of welfare economics says that if the technical assumptions hold, then any pareto optimum can be achieved by a competitive equilibrium with a suitable redistribution of resources. The pareto optimum that maximizes social welfare is the bliss point on the utility-possibilities frontier, and society can get there with lump-sum redistributions that satisfy the interpersonal equity conditions.

This result may seem relatively unimportant, as few markets are perfectly competitive and many of the technical assumptions are frequently violated. Actual economies operate under, not on, their utility-possibilities frontiers. The result is actually quite powerful, however, at least in principle. Our subsequent analysis will show that if the government has enough policy tools at its disposal to restore pareto optimality when faced with market imperfections and violations of the technical assumptions and if it can redistribute resources in a lump-sum fashion, then it should use the lump-sum redistributions to satisfy the interpersonal equity conditions. This is a much stronger statement and suggests the vital role of the interpersonal equity conditions in normative public sector theory. Conversely, if the government does not act to satisfy the interpersonal equity conditions, then it should not necessarily try to achieve the pareto-optimal conditions either. The interpersonal equity conditions and the pareto-optimal conditions go hand in hand in maximizing social welfare; they are both first-order conditions for a social welfare maximum in a first-best policy environment.

The requisite policy tools may not exist to reach the bliss point. Governments may neither be able to restore pareto optimality nor redistribute lump sum. If so, then the policy environment is second-best, and the interpersonal equity conditions no longer provide a theory of optimal income distribution and redistribution. We turn to this important point in Chapter 3.

REFERENCES

Arrow, K., 1983. Contributions to welfare economics. In: Brown, E.C., Solow, R. (Eds.), Paul Samuelson and Modern Economic Theory. McGraw-Hill, New York.

Arrow, K., Sen, A., Suzumura, K. (Eds.), 2002. A Handbook of Social Choice and Welfare Economics, Volumes 1 and 2. Elsevier, North Holland (Vol. 1) and 2010 (Vol. 2).

Bator, F.M., March 1957. Simple analytics of welfare maximization. American Economic Review Vol. 47 (1), 22–60.

Bergson, A., 1938. A reformulation of certain aspects of welfare economics. Quarterly Journal of Economics Vol. 52 (2), 310–334.

Leontief, W., 1966. Essays in Economics: Theories and Theorizing. Oxford University Press, New York.

Mueller, D., June 1976. Public choice: a survey. Journal of Economic Literature Vol. 14 (2), 395–433.

Roemer, J., 1996. Theories of Distributive Justice. Harvard University Press, Cambridge, MA.

Samuelson, P.A., November 1954. The pure theory of public expenditure. Review of Economics and Statistics Vol. 36 (4), 387–389.

Samuelson, P.A., November 1955. Diagrammatic Exposition of a theory of public expenditure. Review of Economics and Statistics Vol. 37 (4), 350–356.

Samuelson, P.A., 1965. Foundations of Economic Analysis. Atheneum Publishers, New York.

Samuelson, P.A., 1981. Bergsonian welfare economics. In: Rosefielde, S. (Ed.), Economic Welfare and the Economics of Soviet Socialism: Essays in Honor of Abram Bergson. Cambridge University Press, New York.

Chapter 3

First-Best and Second-Best Analyses and the Political Economy of Public Sector Economics

Chapter Outline

Chapter 3 concludes our introduction to normative public sector economics with a discussion of two issues. One is the distinction between first-best and second-best analyses. The other is the political economy of public sector theory, centered on the social welfare function and Arrow's impossibility theorem. The social welfare function is the one indispensable political element in normative mainstream public sector models.

LUMP-SUM REDISTRIBUTIONS AND PUBLIC SECTOR THEORY

Are lump-sum redistributions a feasible policy tool for the government? This may appear to be a relatively uninteresting question. One is tempted to answer: "Probably not, but even if they are feasible, it hardly matters because few governments use lump-sum taxes and transfers. For instance, no major US tax or transfer program is lump sum." All this is true, yet it is hard to imagine a more important question for normative public sector theory. The answer has a dramatic impact on all normative policy prescriptions in every area of public sector analysis, whether they are directed at distributional or allocational problems. In public sector theory, lump-sum redistributions stand at the border between first-best and second-best analyses.

The issue is not so much the existence of lump-sum redistributions. Lump-sum tax and transfer programs are easy enough to describe. Poll taxes have occasionally been used as revenue sources and they are certainly lump sum from an economic perspective. On the transfer side, many countries have instituted per-person demogrants (e.g., Canada, which provides a grant to all the elderly). The United States allows a personal exemption for each dependent child under the federal personal income tax. It might be argued that decisions on family size are essentially economic and would influence the amount of transfer received. If so, then tax exemptions and demogrants to children are not strictly lump sum, although the legislation could be drafted such that only children already living at the time of passage would receive the transfers.

The mere existence of lump-sum taxes and transfers is not enough, however, to render them feasible policy tools

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00003-0

in the pursuit of equity. The lump-sum taxes and transfers must be flexible enough so that they can be designed to satisfy the interpersonal equity conditions for social welfare maximization, and this is a very tall order indeed. To be effective, the taxes and transfers would almost certainly have to be related to consumption or income or wealth in order to distinguish the haves from the have-nots, but then it is doubtful that they would be lump sum.

Income taxes were thought to be essentially lump sum before 1970, because empirical research had been unable to discover any relationship between income tax rates and either work effort or saving. Research since then, employing detailed micro data sets and sophisticated microeconometric techniques, suggests that labor supply does respond to changes in after-tax wages, certainly the female labor supply. The evidence on saving behavior is more mixed, but saving also appears to respond somewhat to changes in after-tax rates of return.[1] In any event, no one today believes that income-based taxes and transfers are lump sum. Therefore, the assumption that the government can pursue an optimal lump-sum redistribution policy is heroic in the extreme. Nonetheless, public sector economists have been quite willing to employ the assumption of optimal lump-sum redistributions to analyze allocational policy questions in a first-best framework.

FIRST-BEST ANALYSIS

First-best analysis means that the government has a sufficient set of policy tools for whatever problems may exist to restore the economy to the bliss point on its first-best utility-possibilities frontier. By the "first-best" utility-possibilities frontier, we mean the locus of pareto-optimal allocations constrained only by three fundamentals of any economy: individual preferences, production technologies, and market clearance.[2]

The required set of policy tools is broad indeed. If the analysis occurs within the context of a market economy, it is understood either that all markets are perfectly competitive or that the government can adjust behavior in noncompetitive markets to generate the perfectly competitive results. Faced with a breakdown in one of the technical assumptions discussed in Chapter 1, the government must be able to respond with a policy that restores first-best pareto optimality. As we shall discover in Part II, the required policy responses may be exceedingly complex, enough so that they have little hope of practical application. Finally, the government must employ optimal lump-sum redistributions to equalize social marginal utilities of consumption (income) at the first-best bliss point.

The Two Dichotomies in First-Best Models

What is the attraction of first-best analysis, given its stringent and unrealistic assumptions? The answer is that first-best analysis is really the only way to analyze the particular allocation problems caused by breakdowns in the technical assumptions and market imperfections in and of themselves. Consider, first, the role of lump-sum redistributions in this regard.

If lump-sum redistributions are feasible, then the problem of social welfare maximization dichotomizes into separate efficiency and distributional problems, exactly as the model in Chapter 2 dichotomized into the pareto-optimal and interpersonal equity conditions. The intuition for why this is so can be seen in terms of concepts already developed.

Suppose one of the technical assumptions in Chapter 1 fails to hold, for example, there exists a consumer eternality, meaning that at least one person's utility depends on the goods demanded and/or factors supplied by some other consumer(s). Suppose, further, that the government consists of an allocation branch charged with designing policies to correct for allocational problems such as externalities and a distributional branch charged with creating an optimal distribution of income.[3] If lump-sum redistributions are possible, the allocation branch can ignore the existence of a social welfare function and analyze the externality in the context of the first general equilibrium model presented in Chapter 2, the model in which one consumer's utility is maximized subject to the constraints of all other utilities held constant (and production and market clearance). This model is specifically designed to find the set of pareto-optimal allocations consistent with society's first-best utility-possibilities frontier given the presence of an externality or any other imperfection. All relevant structural elements of the policy necessary to correct for the externality follow

1. For an excellent review of the early empirical studies on labor supply and savings elasticities, see Boskin (1976). The Tax Reform Act of 1986 led to renewed interest in these elasticities. See Auerbach and Slemrod (1997).

2. If some factors or production are supplied in absolutely fixed amounts, they, too, act as constraints on the set of attainable utility possibilities. Recall that the general equilibrium model of Chapter 2 assumes variable factor supplies so that, formally, consumers' disutility from supplying factors enters as an argument of the social welfare objective function rather than as a constraint.

3. Richard Musgrave, the dean of living public sector economists, long ago proposed the useful fiction of government policy emanating from three distinct branches of government, an allocation branch, a distribution branch, and a stabilization branch. The allocation branch was dedicated to pursuing efficiency, the distribution branch to pursuing equity, and the stabilization branch to pursuing long-run economic growth and the smoothing of the business cycle. One difficulty with Musgrave's fiction is the extent to which the three branches can design policies independently from one another. They can operate independently in a first-best environment, but not in a second-best environment. See Musgrave (1959).

directly from the first-order conditions of this model. The allocational branch does not have to worry about social welfare. It knows that the distributional agency is simultaneously designing policies to ensure that social marginal utilities are equalized along the first-best utility-possibilities frontier in accordance with the interpersonal equity conditions. Therefore, it knows that any unwanted distributional consequences of its allocational policies are being fully offset by the distribution branch.

Suppose, instead, that a single superagency concerns itself with both the externality and the original nonoptimal income distribution and develops a full model of social welfare maximization to analyze these two problems simultaneously. Since the first-order conditions of the model dichotomize, this agency would discover one set of pareto-optimal conditions that do not involve the social welfare rankings and one set of interpersonal equity conditions that equalize all social marginal utilities of income (or of one good or factor). These conditions would be identical with those developed independently by the separate allocation and distribution branches. Since the pareto-optimal conditions contain no social welfare terms, they must generate the first-best utility-possibilities frontier. No other result is consistent with social welfare maximization under first-best assumptions. Similarly, the interpersonal equity conditions must be identical to those developed by the independent distribution agency. Only one distribution is consistent with the bliss point on the first-best utility possibilities frontier under the assumptions used throughout the text.

The two independent branches would have to coordinate their efforts. Since an economy is an interdependent system, all allocational decisions have distributional consequences, and vice versa. Consequently, the allocation branch cannot finally set its policies until it knows what the distributional branch has done or is about to do, and vice versa. Continuing with the externality example, suppose the externality is a "bad" such as pollution. Moreover, suppose the correct policy takes the form of a tax on the polluters (a reasonable supposition, as we shall discover in Chapter 6). By following the independent modeling process described above, the allocation branch can determine all the relevant design characteristics of the tax, such as what should be taxed and what parameters in the economy affect the level of the tax rates, but the exact level of the tax rate cannot be determined. The criterion of pareto optimality admits to an infinity of allocations, all of those on the utility-possibilities frontier. In this example, each allocation has one particular tax rate associated with it, so that the final tax rate cannot be announced until the distribution branch announces its optimal redistributional policy, thereby selecting the allocation consistent with the bliss point.

Turning the example around, the interpersonal equity conditions tell the distributional agency all the relevant *design characteristics* of the optimal lump-sum redistributions, but the exact *levels* of all individual taxes and transfers depend in part upon the gains and losses occasioned by the pollution tax. Thus, while it is possible analytically to distinguish between the design of allocational policies and the design of distributional policies, as first-best analysis does, the exact policies to be followed must be simultaneously determined. In formal terms, the pareto-optimal and interpersonal equity conditions are both necessary conditions for social welfare maximization. They must be solved simultaneously to determine a social welfare maximum.

Despite the ultimate interdependence of allocational and distributional policies, the first-best literature on public expenditure theory typically analyzes only efficiency problems inherent in the breakdown of the technical assumptions (or of market imperfections), ignoring completely the question of distributive equity. The analysis generally proceeds along the following lines. First, the pareto-optimal conditions are derived, given that one of the technical assumptions fails. Then policies are described that generate the pareto-optimal conditions, given the assumption that consumers and firms operate within a perfectly competitive market economy. Perfect competition is the only market environment consistent with first-best analysis. The assumption of perfect competition naturally leads to two further questions:

1. What allocation of resources would the competitive market generate in the absence of government intervention?
2. Can the government restore first-best pareto optimality while maintaining existing competitive markets, or is a complete government takeover of some activity absolutely necessary? That is, can the policy be decentralized?

Distributional issues are ignored in the first-best literature not because they are unimportant but rather because they are relatively uninteresting. As noted in the conclusion to Chapter 2, having said that the government should redistribute lump sum to satisfy the interpersonal equity conditions necessary for social welfare maximization, there is little else to say. A breakdown in one of the technical assumptions may alter the precise form of the interpersonal conditions somewhat, but they still have the interpretation that one good (or factor) should be redistributed lump sum to equalize the social marginal utilities of that good (or factor).

In contrast, the pareto-optimal conditions often change substantially when the technical assumptions fail, both in their form and their interpretation. Small wonder, then, that first-best analysis tends to emphasize these conditions and often relegates the interpersonal equity conditions to a footnote, if they are mentioned at all. Knowing that the first-order conditions of a full model of social welfare maximization dichotomize, there is no need to use the full model. A simple

model highlighting the first-best pareto-optimal conditions for the allocational problem at hand is sufficient.

The first-best analysis in Part II of the text is careful, however, to use full models of social welfare maximization when analyzing allocational problems. Keeping the social welfare function in the models serves to emphasize the importance of lump-sum redistributions to all first-best policy analysis.

First-best models have a highly useful second dichotomy property besides the dichotomy between the pareto-optimal and interpersonal equity conditions. The pareto-optimal conditions themselves dichotomize. A breakdown in one of the technical assumptions or a market imperfection alters the pareto-optimal conditions for those goods and factors directly affected but leaves unchanged the form of the pareto-optimal conditions of all the unaffected goods and factors. For example, suppose a competitive market satisfies the pareto-optimal condition for the allocation of some good, with price equal to marginal cost. Price equal to marginal cost continues to be the pareto-optimal pricing rule for that good even if other markets contain externalities or exhibit decreasing cost production, so long as the policy environment is first best. The government's response to the market failure can stay focused on the source of the market failure.

To summarize, the double dichotomy of distributional and allocational problems under first-best assumptions makes first-best analysis especially attractive for the *ceteris paribus* analysis of policy issues. An allocational problem associated with a particular economic activity can be isolated from distributional considerations and from all the other conditions within the economy that are required for pareto optimality. This property justifies the use of very simple general equilibrium models that focus exclusively on one source of market failure and describe the rest of the economy by means of a single composite commodity that is assumed to be marketed competitively. Assuming a first-best policy environment is a tremendous analytical convenience.

SECOND-BEST ANALYSIS

Suppose, realistically, that lump-sum taxes and transfers are not available to the government, at least not with sufficient flexibility to generate the interpersonal equity conditions of the standard model. This changes the analysis rather drastically. To see why, consider two government policy strategies in the context of a market economy, one designed to produce distributive equity, the other designed to restore first-best pareto optimality.

Constrained Social Welfare Maximization

Suppose that the government chooses to redistribute income until social marginal utilities are equalized by using taxes and transfers that are not lump sum.[4] The redistribution necessarily introduces distortions into the economy because some consumers and/or producers now face different prices for the same goods and/or factors. Since consumers and producers equate relative prices to their marginal rates of substitution and transformation, respectively, and since pareto optimality requires that the marginal rate of substitution (MRS) equals the marginal rate of transformation, some of the pareto-optimal conditions no longer hold. The redistribution forces the economy beneath its first-best utility-possibilities frontier.

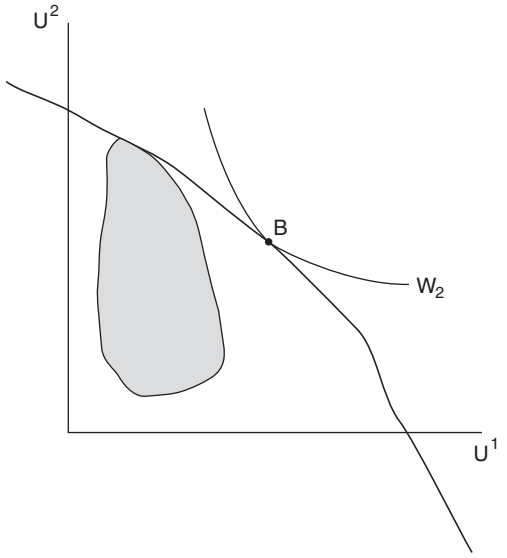

FIGURE 3.1

Suppose instead that the government focuses only on allocational problems and chooses allocational policies designed to bring society to the first-best utility-possibilities frontier.[5] Without simultaneously employing lump-sum redistributions, however, the economy would not be at the bliss point, in general. The government may actually choose some policy mix designed to move the economy somewhat closer to full pareto optimality, and somewhat closer to distributive equity, but the point remains that removing the possibility of feasible lump-sum redistributions restricts the set of solutions available to the government, for example, to the shaded portion in Fig. 3.1. The viable allocations and distributions may or may not include points on the first-best utility-possibilities frontier, but, importantly, they definitely exclude the bliss point, point B. The policy problem now becomes one of finding the best policy option within this restricted set of opportunities. As such it is part of *second-best analysis*, defined as the analysis of optimal public sector policy given that the

4. Assume it is possible to equalize social marginal utilities without lump-sum redistributions. It may not be, given the available policy tools.
5. Again, assume this is possible.

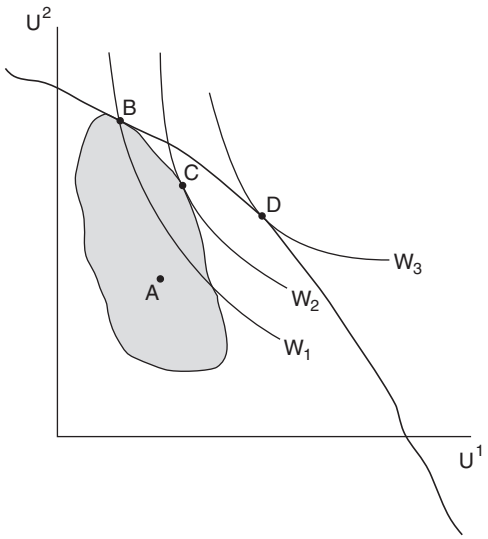

FIGURE 3.2

bliss point on the first-best utility possibilities frontier is unattainable.

One immediate implication of second-best public expenditure analysis is that government policy should not necessarily try to keep society on its first-best utility-possibility frontier. Points other than those on the first-best frontier may yield greater social welfare within the restricted set of policy alternatives. To see this, refer to Fig. 3.2.

Suppose society is initially at point *A* in Fig. 3.2, possibly because one of the technical assumptions of Chapter 1 has failed and the competitive market is therefore generating the wrong allocation of resources. If lump-sum redistributions were feasible and the world were otherwise first best, the government should design policies to restore full pareto optimality and redistribute lump sum to achieve the bliss point, point *D* in the figure. In a second-best environment, without the ability to redistribute lump sum, the policy option that brings society to point *B* on the first-best frontier is dominated by another option that keeps society below the frontier, point *C*. Society's goal is still the maximization of social welfare, reaching the highest possible social welfare indifference curve. Point *C* is the maximum attainable level of social welfare given the restricted set of available options. Point *B* is pareto efficient, point *C* is not, but the superior distributional attributes of point *C* prove decisive. In a second-best environment, then, society's efficiency and equity norms are completely interrelated. They cannot be pursued with separate policy tools, unlike in a first-best policy environment.

Figure 3.2 highlights the way we defined second-best policy analysis, as the inability to attain the first-best bliss point. An equivalent definition is the analysis of optimal public sector policy given that additional constraints have been added to the first-best framework of social welfare maximization. The addition of a single binding constraint on the baseline model of Chapter 2 renders the first-best bliss point unattainable, establishing the equivalence of the two definitions of second-best analysis.

As noted above, the first-best constraints consist solely of the fundamental economic constraints of any economy: the production technologies and market clearance equations.[6] The additional constraints are typically either restrictions on the permissible set of government policy tools or maintained imperfections in the market economy. The form of the additional constraint does not matter. Any single additional binding constraint, or any combination of these constraints, renders the analysis second best.

We have been careful throughout this section to refer to the "first-best" utility-possibilities frontier, or the "first-best" bliss point. Given the existence of additional constraints, it is always possible to derive a new utility-possibilities frontier and a new bliss point corresponding to the restricted set of feasible allocations. In terms of Fig. 3.2, these would correspond to the outer boundary of the shaded portion and point *C*, respectively. The government can still be thought of as pursuing the distributionally best allocation among all possible pareto-optimal allocations along the restricted frontier, exactly as in first-best analysis. Although this is technically correct, it tends to obscure the important differences between first-best and second-best analysis, differences that can best be seen in terms of the attainable first-best allocations.

The Most Common Policy and Market Constraints

The inability to redistribute lump sum is merely one of a large number of possible constraints on the feasible set of government policy tools. It is often one of the constraints chosen, because second-best analysis is an attempt to develop normative policy rules in more realistic policy environments, and denying the government feasible lump-sum redistributions is an obvious step toward realism. Generalizing beyond this is more difficult. The second-best literature has considered an enormous variety of additional constraints on available policy tools. This is hardly surprising, since the set of potential constraints is virtually limitless, given political realities and the staggering complexity of actual market economies. It is fair to say, however, that four kinds of policy restrictions have been most commonly employed in the public sector literature (in addition to restrictions on lump-sum redistributions): the use of distorting taxes and transfers, the existence of legislated budget constraints on individual government

6. And fixed factor supplies, if relevant, see footnote 2.

agencies or on the government as a whole, the drafting of resources or the offer of certain government services free of charge (or at prices below marginal cost), and asymmetric information in the form of private information about individuals that the government cannot know (at least not without bearing some costs to monitor the individual).

Distorting Taxes and Transfers

Almost no major tax or transfer programs are lump sum. Actual taxes are either *ad valorem* (percentage of price) or per-unit taxes on buyers or sellers of goods and factors, including sales and excise taxes on goods and services, income and payroll taxes on factors of production, and various kinds of wealth or property taxes. In addition, tax rates on income and wealth are often graduated, increasing with income (wealth). All these taxes force buyers and sellers to face different prices in the same markets and are thereby distorting. Most of the major transfer programs condition the transfers on consumption or income, which makes them distorting as well. Therefore, models that analyze actual distorting taxes and transfers directly or assume that distorting taxes are being used to finance public expenditures are necessarily second best. In contrast, taxes used to solve problems such as externalities in a first-best environment promote social welfare. They cannot be distorting in the sense of generating welfare costs.

Analysis of the welfare costs of distorting taxes and transfers has a long history dating back to the very beginnings of public sector economics. The discipline was named public finance until about 50 years ago because the emphasis was more on tax policy than on expenditure policy. Public finance economists studied a number of issues related to the efficiency of taxes that have no meaning in a first-best environment, including the following: If the government must raise revenue using a single distorting tax, such as a particular sales or income tax, what are the efficiency costs to society? Are some taxes less costly (that is, less distorting) than others per dollar of revenue collected? If the government is free to vary a wide set of distorting taxes, what pattern of tax rates minimizes the resulting distortions while raising a required amount of revenue? The allocational theory of taxation has always been a second-best analysis. The main change in tax theory over the past 30 years is that general equilibrium modeling techniques have increasingly replaced partial equilibrium analysis in studying these issues. It is not that assumptions with respect to tax instruments have become more realistic.

The assumption that governments use distorting taxes to finance government expenditures has become commonplace over the past 50 years, and it has had a monumental impact on public expenditure theory. The problems being analyzed are the same as in the older first-best analysis—principally externalities and decreasing cost production—but the second-best optimal policy prescriptions are often dramatically different from their first-best counterparts.

Fixed Budget Constraints

Legislatures usually impose budgetary ceilings on individual government agencies that can be exceeded only by means of special supplemental appropriations. Frequently, the budgets of entire governments are limited as well. In the United States, for instance, many state and local government administrations are required to submit annually balanced operating budgets. Even without this requirement, most state and local governments cannot routinely borrow in the national capital markets to cover annual operating deficits without threatening their credit ratings. Only the federal government enjoys this privilege.

Imposing either agency-by-agency or overall budget constraints is generally not a first-best strategy. Only by chance would legislators set budgets at the expenditure levels consistent with a first-best allocation of resources. Thus, as a further step toward reality, public sector economists have incorporated legislated budget constraints into their models to see how they affect traditional first-best policy rules. Once again, the new second-best policy prescriptions are often quite different from their first-best counterparts.

Drafting Resources or Giving away Goods

All scarce goods and factors have marginal opportunity costs associated with them. Their prices would reflect these marginal opportunity costs in a first-best world, but governments sometimes choose to set prices well below opportunity costs, often at zero. The military draft is one example on the factor side; citizens are required to serve and many are paid below their market wages. On the goods and services side, governments in the United States often follow an average cost pricing strategy when they do charge for public services. Sometimes they just give public services away, such as the side benefits of hydroelectric projects in the form of flood control protection to homes and irrigation of farmland. These self-imposed government pricing constraints have often been the focus of second-best analysis.

Maintained Monopoly Power

Market imperfections would render the first-best bliss point unattainable even if government policy tools were not restricted in any of the ways described above. One example is monopoly power. Price does not equal marginal cost in markets with monopoly power, so that the pareto-optimal conditions do not hold for these goods and services. Monopoly power could be viewed as a restriction on government

policy in the sense that the government is unable to correct the imperfection despite the existence of policies that would do so. In any event, any maintained market imperfection such as monopoly power implies a second-best environment, and the first-best policy rules of public expenditure theory may not be optimal.

Asymmetric or Private Information

Another pervasive market imperfection is asymmetric or private information. We described in Chapter 1 the various ways in which private information leads to a call for government intervention, e.g., to establish a legal system and bureaus of standards and to provide public insurance. We also noted the difficulties it poses for the government's responses to all problems under the government-as-agent ground rules. Recall that the problem of private information is not limited to allocational issues. Private information is a decided handicap to a government interested in redistributing purchasing power in accordance with the interpersonal equity conditions of first-best theory. Redistributional policies can hardly be effective if people can hide their incomes from the government. Suffice it to add here that second-best analysis now commonly includes private information as one of the constraints that prevents government policy from attaining the first-best bliss point.

Further Implications of Second-Best Modeling

Two further distinctions between first-best and second-best modeling are worth emphasizing in these introductory comments, both resulting from the feature that second-best general equilibrium models are basically first-best models modified by the addition of one or more constraints.

The Scope of Government Intervention

As noted earlier, the first-order conditions of first-best models dichotomize in two ways that are especially convenient for *ceteris paribus* policy analysis. Second-best models typically do not dichotomize in either way. As a general rule, all the necessary first-order conditions of a second-best model contain both efficiency and equity considerations, especially if lump-sum distributions are not permitted. This is simply the formal counterpart to a point demonstrated by Fig. 3.2, that the efficiency and equity norms are directly interrelated when the first-best bliss point is unattainable.

This property of second-best models has been especially disheartening for normative analysis because it further limits the government's ability to honor the principle to consumer sovereignty. In a first-best environment, the demand (factor supply) content of all allocational decision rules derives solely from individual's preferences, usually their marginal rates of substitution. The social welfare rankings influence allocational decision rules only indirectly in the sense that any redistribution can be expected to shift aggregate demands. Thus, consumer sovereignty guides the government's intervention into the market economy when addressing allocational problems. In a second-best environment, however, the allocational decision rules contain the social welfare rankings as well as terms representing individuals' preferences, so that consumer sovereignty must be partially overridden even in allocational decision making. This property of second-best analysis is doubly disturbing, since there is nowhere near a consensus on what the social welfare rankings should be. It is no longer possible to isolate the uncertainties associated with the social welfare rankings into a single decision on optimal income distribution.

Worse yet, the social welfare terms contaminate all markets in general, even those that first-best analysis would leave entirely in the hands of the competitive market system. This is so because a second-best policy environment generally requires broad intervention of the government into the workings of the market economy, unlike the more limited intrusion of first-best analysis. Government intervention remains justified by market failure, but the intervention is no longer limited to the markets containing the failures. Policy prescriptions that require broad government intervention are naturally resisted in capitalist societies.

The broader intrusion of the government in a second-best environment follows directly from a famous theorem published in 1956 by Archibald Lipsey and Kelvin Lancaster. They proved that if the first-best pareto-optimal conditions are assumed not to hold for some goods and factors as a maintained hypothesis, then it is generally not optimal to pursue first-best pareto optimality for the other goods and factors. Their article now stands as a classic in public sector economics, and the Lipsey—Lancaster theorem is often referred to as the theorem of the second-best Lipsey and Lancaster (1956).

Lipsey and Lancaster spoke in terms of the pareto-optimal conditions because the model they used to illustrate the theorem did not contain a social welfare function. Nonetheless, their theorem applies to the broader social welfare model as well. If one of the first-order conditions for a first-best social welfare maximum fails to hold because of an added constraint to the model, then the other first-order conditions do not hold either at the constrained second-best welfare optimum, in general.

Interpreting Second-Best Results

Still another discouraging implication of second-best analysis is that second-best allocational decision rules generally do not have clear intuitive interpretations with obvious analogs to free-market principles. First-best

allocational decision rules often do have competitive analogs, because they are usually just simple combinations of consumers' marginal rates of substitution and producers' marginal rates of transformation (marginal rates of technical substitution for factors). Since competitive markets equate price ratios to these margins, a competitive market structure can always be described that would generate first-best pareto-optimal conditions of this type. This result is especially appealing if one believes in competitive markets, consumer sovereignty, and the least possible amount of government interference with the market system. In first-best analysis, the government can often be viewed as an imitator of perfectly competitive behavior in solving allocational problems.

There are two formal reasons why second-best allocational decision rules tend not to have competitive market interpretations. One is that terms from the additional constraints appear in the first-order conditions along with their associated Lagrangian multipliers, and the multipliers are unrelated to standard market concepts. The other has already been noted, that the decision rules generally contain social welfare terms if lump-sum redistributions are forbidden. The social welfare terms certainly have no competitive market analogs.

Model and Policy Sensitivity

A final discouraging property of second-best optimal policy rules is that they tend to be rather sensitive to modifications in constraints or additions of new constraints. This type of model sensitivity is extremely troublesome because the real world is obviously many times more constrained, more imperfect than any analytical model can hope to capture. Second-best analysis can never hope to produce truly definitive government policy rules on anything.

To summarize, second-best public expenditure theory has offered the severest possible challenge to the long-standing first-best orthodoxy in the attempt to make public sector theory more realistic. The second-best rules often bear no clear-cut relationship to their long-standing first-best counterparts. These challenges notwithstanding, the first-best results of public expenditure theory have hardly disappeared. They still dominate undergraduate textbooks on public sector economics and they instruct much actual policy debate. The staying power of first-best analysis is no doubt due to the intuition it provides about allocational and distributional issues and its call for limited government intervention. In contrast, second-best policy rules tend to be resisted as normative policy prescriptions; since they require ethical or distributional judgments associated with the problematic social welfare terms, they tend to call for broad intervention in the economy; and the policy rules are so sensitive to the form and number of constraints. The relative advantages of the first-best results must always be weighed against the blatant unrealism of the first-best models in a policy context.

SIMILARITIES BETWEEN FIRST-BEST AND SECOND-BEST ANALYSES

The numerous differences between first-best and second-best public sector analysis should not obscure the fact that the two approaches are virtually identical in method and philosophy. The challenge to first-best orthodoxy is contained in the first-order conditions of the second-best models. One would certainly not want to minimize the importance of this challenge, since the first-order conditions translate directly into normative policy rules. But second-best analysis hardly represents a methodological or philosophical departure from first-best theory. All it does is attach some additional constraints to the basic first-best neoclassical general equilibrium model in an attempt to be more realistic. This is not revolutionary. For instance, second-best analysis retains the fundamental notion that the government is interested in social welfare maximization, with social welfare indexed by means of an individualistic Bergson–Samuelson social welfare function. In principle, then, second-best analysis honors consumer sovereignty to the same degree as first-best analysis, even though its results are less clear cut in this regard. Furthermore, second-best research has generally remained closely allied with the competitive market system, so much so that the following standard competitive market assumptions are commonplace in second-best models:

1. Consumers maximize utility subject to a budget constraint and have no control over any prices.
2. Private sector producers are decentralized price-taking profit maximizers such that goods prices equal marginal costs and factor prices equal the values of marginal products.
3. If there is government production, the government buys and sells factors and outputs at the competitively determined private sector producer prices. This is often true even if the second-best decision rules imply that a different set of "shadow" prices should be used to determine the optimal level of government production.

There are two reasons why second-best analysis has emphasized competitive market behavior. The first turns on the *ceteris paribus* condition. Exploring the effects of particular market imperfections or policy restrictions on first-best public sector decision rules requires introducing them as constraints one at a time into an otherwise first-best model. If the analysis proceeds within the context of a market economy, this means that the parts of the market economy not specifically analyzed must be assumed to be competitive. As a result, second-best analysis to date has

been much closer to a first-best perfectly competitive market environment than to highly imperfect real-world market economies. It is at best a small, hesitant step toward reality.

The second reason is also a matter of analytical convenience. The competitive market assumptions permit flexibility in model building, a feature that second-best analysis has frequently exploited. As noted in Chapter 2, general equilibrium models can always be defined in terms of quantities of goods and factors, with the economy viewed as being under the control of a social planner. The model developed in that chapter served as an example. General equilibrium models can also be expressed directly in terms of prices by incorporating specific assumptions about market structure and behavior, and the competitive assumptions happen to be the easiest ones to employ. For many second-best problems, the price specification has proven to be the most direct analytical approach.

To gain some preliminary intuition why this is so, consider the common second-best policy restriction that the government must use distorting taxation. As noted above, taxes distort by driving a wedge between the prices faced by different economic agents operating in the same market. If the general equilibrium model is already defined in terms of prices, the gross and net of tax prices (and the tax itself) can be incorporated directly into the model. Furthermore, all of the interesting allocational and distributional implications of the tax follow directly from the first-order conditions of the price/market model. Proceeding in this way turns out to be far more convenient than beginning with a quantity model and reworking the first-order conditions using standard market assumptions to capture the effects of the taxes.

In summary, the transition to the second-best analysis in Part III of the text from the first-best analysis in Part II is fairly easy and straightforward.

THE POLITICAL ECONOMY OF THE SOCIAL WELFARE FUNCTION

The social welfare function is central to mainstream normative public sector theory. It is the only indispensable political element of the theory and it serves two critical analytical purposes: It describes society's views on distributive justice, and it selects the one efficient allocation that maximizes social welfare from the infinity of possible efficient allocations. At the same time it is a highly problematic concept because of the limitations noted in Chapter 2. The two most serious limitations for the normative theory are the lack of a consensus on what the social welfare function should be and the difficulties that a democratic society may have in formulating a consistent social welfare function. Each of these limitations deserves some discussion.

The Form of the Social Welfare Function: From Utilitarian to Rawlsian

Neither economists nor anyone else have been able to agree on what the appropriate end-results ethical rankings of individuals should be, that is, what form the social welfare function should take. The only consensus that has emerged in the economic literature is on the reasonable limits of the ethical rankings. Most economists agree that the ethical spectrum should be bounded by utilitarianism at one end and Rawlsianism at the other end. Utilitarianism implies complete indifference to the distribution; Rawlsianism implies the greatest degree of equality.

Utiltarianism

The utilitarian view reached its height of popularity among social philosophers and political economists in late 1700s and early 1800s under the leadership of Jeremy Bentham. Bentham and his followers argued that the goal of society should be to maximize aggregate happiness or satisfaction. Their view implies that social welfare is the sum of the individuals' utility functions:

$$W = \sum_{h=1}^{H} U^h \tag{3.1}$$

where W is the utilitarian or Benthamite social welfare function. Its social welfare indifference curves for any two individuals are 45° straight lines as pictured in Fig. 3.3.

One appealing feature of utilitarianism is its adherence to the ethical principle of impersonality, that all people should have equal ethical weight. The ethical weights of a social welfare function are the marginal social welfare terms $\partial W/\partial U^h$, which are all equal to one under utilitarianism. Societies do not always honor the impersonality principle, however. Affirmative action in the United States

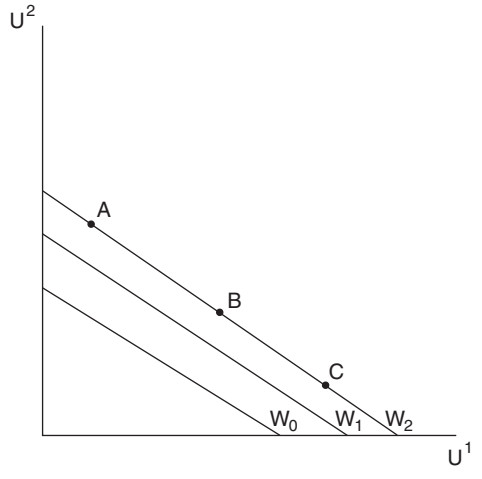

FIGURE 3.3

is a counterexample, with its other-things-equal preference for women and minorities in hiring. Yet, some compelling justification usually lies behind the violations of impersonality in liberal societies, such as the unfair handicaps resulting from current and past discrimination in the case of affirmative action. Another appealing feature of utilitarianism is that it honors the pareto principle. Social welfare increases (decreases) if one person is made better off (worse off) and no other person's utility changes.

These appealing features are more than counterbalanced for most economists by utilitarianism's complete indifference to the distribution of well-being. Points A, B, and C on social welfare indifference curve W_2 in the figure all yield the same amount of social welfare. Societies are never indifferent to such extremes in the distribution, however, and most people are not either.[7]

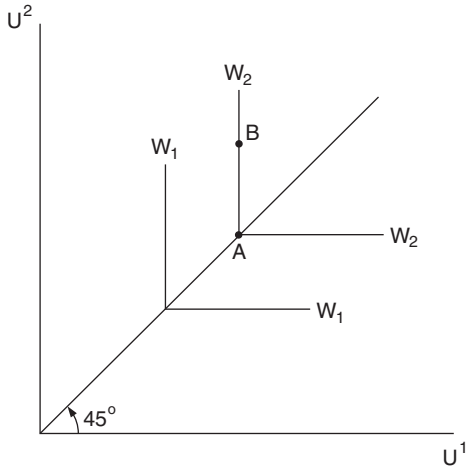

FIGURE 3.4

Rawlsianism

Rawlsianism is named after the ethical position described by Harvard philosopher John Rawls Rawls (1971). Rawls argues that people have difficulty thinking about end-results equity because they know where they stand in the distribution and have reasonably firm expectations about their future well-being. The only way people can think objectively about distributive justice, according to Rawls, is to assume that they stand behind a veil of ignorance, with no idea at all about their current or future position in the distribution. In other words, people should assume they are truly uncertain about their prospects, unable even to attach probabilities about their possible outcomes. As such, they cannot choose to maximize expected utility, the standard assumption about consumer behavior under uncertainty.

What principles of distributive justice would people adopt in the face of true uncertainty about the distribution? Rawls believed that people would become extremely risk averse and adopt a maximin strategy. They would agree that society should always pursue policies that maximize the well-being of those who are the worst off, based on the possibility that they could be among the worst off at some future date. Rawls' position implies the Min form for the social welfare function:

$$W = \min_h \left(U^1, \ldots, U^h, \ldots, U^H \right) = \min_h \left(U^h \right) \quad (3.2)$$

where W is the Rawlsian social welfare function. Its social welfare indifference curves for any two individuals are right angles from the 45° line of equality as pictured in Fig. 3.4. Movement from point A to B along W_2 in the figure does not increase social welfare because person 1 is now the worst off at point B, and that person's social welfare has not improved relative to point A. Social welfare can only increase from a position of equality if both peoples' utilities increase, because then the worst-off person becomes better off. In addition, equality is the social welfare maximum for any given aggregate level of utility because then the worst-off person has the highest possible utility.

Rawl's veil of ignorance principle when thinking about distributive justice is very appealing to many people. It is a central tenet of public choice theory. As noted in Chapter 1, Buchanan uses it as the justification for why the self-serving framers of the constitution would allow governments to redistribute income. Overall, though, Rawls' position has been rejected by the majority of economists. It is highly problematic from an economic perspective.[8]

To begin with, why should people be so extremely risk adverse in the face of true uncertainty that they favor the maximin strategy? Economists have not been able to develop a consensus theory of behavior under true uncertainty, but maximin is just one of many possible strategies that people might adopt. Also, the maximin strategy has a number of unattractive features. It suggests, for example, that people would forego the possibility of a new situation that makes the worst-off individuals slightly worse off and everyone else substantially better off. A vast majority of people might be willing to accept the new situation on the chance that they would not be among the new worst off. An especially uncomfortable example of this possibility relates to long-run economic growth. Virtually all societies favor economic growth, yet saving for growth is not a maximin strategy in an intergenerational context. The first generation is always the worst-off in a growing economy, and saving for the benefit of future generations makes them even worse off. Still another severe drawback of Rawlsianism is that it

7. An excellent discussion of the pros and cons on utilitarianism from both economic and philosophic perspectives is found in Gordon (1980).

8. See Arrow (1973).

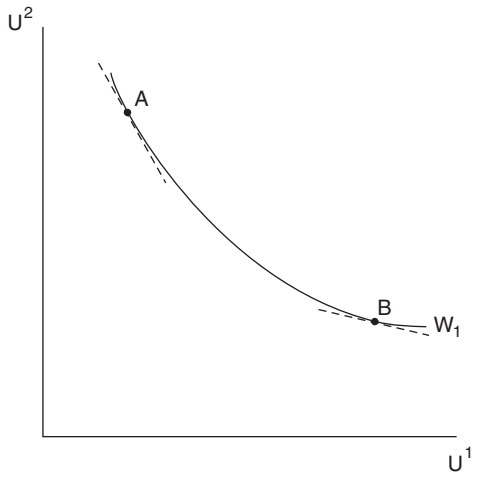

FIGURE 3.5

does not honor the pareto principle, as the move from point A to B in Fig. 3.4 illustrates.

Most economists believe that social welfare indifference curves should have the standard property of diminishing MRS and be convex to the origin, as shown in Fig. 3.5. Compare points A and B. At point A, when person 2 is much the better off, society should be willing to sacrifice more of person 2's utility to make person 1 one-util better off than it would be willing to sacrifice at point B, when person 1 is already much better off. The greater the curvature of the indifference curves the more egalitarian the social welfare function, with the straight-line curves of utilitarianism and the right-angled curves of Ralwsians defining the reasonable extremes.

A Flexible Social Welfare Function

In theoretical work it is often unnecessary to be specific about the form of $W(\ )$. $\partial W/\partial U^h$ is simply understood to represent society's ethical judgment about the social marginal utility of person h, whatever that judgment may be. A specific parameterization of $W(\ )$ is essential, however, if one wishes to test the sensitivity of a normative policy rule to society's social welfare rankings. Many different functional forms appear in the literature, but the most common is one suggested independently by Anthony Atkinson and Martin Feldstein in the early 1970s.[9] It has the advantage of being able to represent the full range of possibilities from the utilitarian to the Rawlsian positions. Define $W(\ )$ as

$$W[U_h()] = \left(\sum_{h=1}^{H} U_h^V\right)^{1/V}, \quad V = [1, \ -\infty] \tag{3.3}$$

where V is a constant reflecting society's aversion to inequality. $V = 1$ implies that W is the straight sum of the individuals' utilities, utilitarianism. At the other extreme, $V \to -\infty$ implies maximizing the utility of the worse-off individual, the Rawlsian maximin criterion.[10] In between the extremes, increasingly larger negative values of V imply increasing aversion to extremes in the distribution of utility.[11] The social welfare indifference curves become ever more convex to the origin. Equation (3.3) is an especially convenient flexible functional form for examining the robustness of policy rules to distributional judgments because the various possibilities are contained in the single parameter V, the aversion to inequality.

9. Atkinson (1973), Feldstein (1973). These specifications assume the government has chosen particular cardinal representations of the U_h on which to base its social welfare judgments.

10. To see that $V \to -\infty$ implies the Rawls maximin criterion, differentiate W with respect to some U_j:

$$\frac{\partial W}{\partial U_j} = \frac{1}{V} \cdot \left(\sum_{h=1}^{H} U_h^V\right)^{(1/V-1)} \cdot V U_j^{(V-1)} \tag{3.3a}$$

$$\frac{\partial W}{\partial U_j} = \frac{U_j^{(V-1)}}{\left(\sum_{h=1}^{H} U_h^V\right)^{\frac{V-1}{V}}} = \left[\frac{U_j}{\left(\sum_{h=1}^{H} U_h^V\right)^{1/V}}\right]^{(V-1)} \tag{3.3b}$$

Dividing numerator and denominator inside the brackets by U_j and rearranging terms yields

$$\frac{\partial W}{\partial U_j} = \frac{1}{\left[\sum_{h=1}^{H} \left(\frac{U_h}{U_j}\right)^V\right]^{(1-1/V)}} \tag{3.3c}$$

Letting $V \to V \to -\infty$ yields

$$\frac{\partial W}{\partial U_j} = \frac{1}{\frac{1}{\left(\frac{U_1}{U_j}\right)^{\infty}} + \ldots + \frac{1}{\left(\frac{U_h}{U_j}\right)^{\infty}} + \ldots + 1 + \ldots + \frac{1}{\left(\frac{U_H}{U_j}\right)^{\infty}}} \tag{3.3d}$$

If U_j is selected such that $U_j < U_h$, $j \neq h$, all variable terms in the denominator of Eqn (3.3d) go to zero in the limit, so that

$$\frac{\partial W}{\partial U_j} = 1 \tag{3.3e}$$

Selecting any other U_j implies that the denominator becomes large without limit. Hence,

$$\frac{\partial W}{\partial U_j} = 0, \quad U_j \neq \min_{(h)} U_h$$

11. By inspection of Eqn (3.3c), the value of $\partial W/\partial U_j$ increases as V becomes increasingly negative, $U_j < U_h$ for all $j \neq h$.

Arrow's Impossibility Theorem

The social sciences ran headlong into a brick wall in 1951 when Kenneth Arrow published his general impossibility theorem. There is no other way to put it. Arrow's theorem is truly devastating to democratic societies.

Arrow was commissioned by the Department of Defense to develop a theory of how democratic societies should make decisions about public goods such as defense. He approached the problem of social decision making in the manner of cooperative game theory: Develop a minimal set of axioms to guide the social decision process that would be acceptable to a democratic society and then determine the implications of those assumptions. Arrow put forth five axioms that he thought a democratic social decision process should possess. He then proved that, in general, no social decision process can simultaneously satisfy all five axioms.

Arrow's theorem does not imply that a democratic society cannot make social decisions. They clearly can, and do. But, it does imply that a democratic society cannot, in general, formulate consistent social decisions under a minimal set of conditions that would be acceptable to it. Arrow's theorem applies to social decisions on any issue, including the attempt to formulate a consistent social welfare function for resolving the problem of distributive justice. All students interested in public sector economics should have at least an intuitive understanding of Arrow's general impossibility theorem. It is considered by many to be the landmark result in twentieth-century political philosophy.

Arrow's Five Axioms

Arrow proposed the following five axioms as reasonable requirements for social decisions in a democratic society:

1. *Universality*: Individuals should be allowed to have any preferences they wish about social outcomes. Democratic societies should not be willing to impose restrictions on individuals' preferences, presuming of course that the preferences are right minded and not destructive.
2. *A complete ordering*: The social decision process must be able to provide a complete ordering of social outcomes for all possible combinations of the individuals' preferences over those outcomes, just as consumers must be able to provide a complete ordering of all possible consumption bundles. One requirement of a complete preference ordering is that it be transitive.
3. *The pareto principle*: The social decision process must honor the pareto principle: If every individual prefers social outcome X to social outcome Y, then society must prefer X to Y. (This is the strong version of the pareto principle.)
4. *The independence of irrelevant alternatives*: Suppose society prefers X to Y, and it also prefers Y to Z. Then individuals change their minds regarding Y and Z and now prefer Z to Y. The change in preference between Z and Y cannot change the preference between X and Y. Z is considered an irrelevant alternative in the choice between X and Y.

This is the least intuitive of Arrow's assumptions, but it is sensible for a democratic decision process. One huge advantage is that it conserves on information in decision making. Without this assumption, the ranking of two alternatives may depend on the rankings of all other alternatives, which can become unwieldy for the decision-making process. Also, the assumption sharply reduces the possibilities for strategic behavior. For example, suppose 10 possible outcomes are under consideration and individuals are asked to rank order each one from 1 to 10. The winning outcome is the one with the highest total score. Suppose one person prefers Y first and X second but is afraid that X will win. That person has an incentive to falsely rank X last to boost Y's chances of winning. The independence assumption rules out such behavior. Suffice it to say that economists have tried to eliminate or replace this assumption without much success in improving the social decision process.

5. *Nondictatorship*: The rankings made by the social decision process cannot always be the same as the ranking of one particular person no matter what the preferences of the other people are. If this were so, the one individual is effectively a dictator.

The nature of the proof is that all five axioms cannot hold simultaneously, in general. The usual way of presenting the proof is to assume axioms one through four hold and then show that these four assumptions imply that one person is a dictator.

To gain an intuition for why one person inevitably becomes a dictator, consider a simple two-person example in which each person has preferences over three social outcomes, X, Y, and Z. There are 36 possible preference pairings over which society must make a choice. To begin, consider the ranking X P Y P Z for person 1 (first column below), paired with all six possible rankings for person 2 (second column):

XX	XX	XY	XY	XZ	XZ
YY	YZ	YX	YZ	YX	YY
ZZ	ZY	ZZ	ZX	ZY	ZX

If the preferences are as in column 1, then society chooses X SP Y SP Z by the pareto principle—they both agree (SP means the first variable is socially preferred over the next). The first disagreement occurs in column 2, between Y and Z. Suppose society chooses in favor of person 1, so that Y SP Z when the two disagree. Having decided

this one time in favor of person 1, society must favor person 1 forever after when the two disagree. The way to show this is to select pairs of preferences such that they agree on one ranking and disagree on the other two, but society has already settled one of the disagreements. The universality assumption (U) allows us to consider pairings in this manner, because every possible pair of preferences must yield a consistent social decision. Then the pareto principle (PP) and transitivity (T) settle the remaining disagreement in favor of person 1.

To see how this works, look at the fifth column. The two agree on X vs. Y, and disagree on Y vs. Z and X vs. Z. Society must decide that X SP Y because the two agree (PP). Also, from the second column, society ranks Y SP Z whenever person 1 says Y P Z and person 2 says Z P Y, as here. But, if X SP Y and Y SP Z, then by transitivity X SP Z. Therefore, society's rankings are the same as those of person 1.

Next we need to determine what happens when the two people disagree on the ranking of X and Y. Suppose that person 1 says X P Y and person 2 says Y P X. To see that person 1 prevails, select the following pairing with one agreement and two disagreements (with the preferences of person 1 in the first column, as always):

XZ
ZY
YX

They agree on Z and Y, so that Z SP Y (PP). Also, when person 1 says X P Z, and person 2 says Z P X, we have seen that X SP Z. Therefore, X SP Z and Z SP Y imply X SP Y (T). Person 1 wins again.

To complete the possibilities, reverse the order of the disagreements. Suppose person 1 says Z P Y and person 2 says Y P Z, the opposite of the first disagreement above which we assumed was decided in favor of person 1. This time select the pair:

ZY
XZ
YX

They agree on Z and X, so Z SP X (PP). Also, when person 1 says X P Y and person 2 says Y P X, we have seen that X SP Y. Therefore, Z SP X and X SP Y imply Z SP Y (T). Person 1 wins again.

Next suppose that person 1 says Z P X and person 2 says X P Z. This time select the pair:

ZY
YX
XZ

They agree on Y and X, so Y SP X (PP). Also, when person 1 says Z P Y and person 2 says Y P Z, we have seen that Z SP Y. Therefore, Z SP Y and Y SP X imply Z SP X (T). Person 1 wins again.

Finally, suppose person 1 says that Y P X and person 2 says that X P Y. This time select the pair:

YX
ZY
XZ

They agree on Y and Z, so that Y SP Z (PP). Also, when person 1 says Z P X and person 2 says X P Z, we have seen that Z SP X. Therefore, Y SP Z and Z SP X imply Y SP X (T). Person 1 wins again.

Person 1 wins all possible disagreements over the pairs of outcomes and is therefore said to be decisive, a dictator, over all pairs of preferences involving X, Y, and Z. (Verify that person 1 must win the remaining pairings that we did not consider in the row of six pairings above, columns 3, 4, and 6. Also, verify that if the first disagreement above is decided in favor of person 2, then person 2 would be the dictator, using the same method of combining one ranking on which they agree and two on which they disagree.) Finally, note that the independence of irrelevant alternatives has been used implicitly in the examples. When deciding on any two outcomes, the position of the third outcome within each person's rankings is irrelevant to the social decision on the two outcomes.

Next, add a new outcome to the list, say W. If person 1 is decisive over all pairs of preferences involving X, Y, and Z, then person 1 must also be decisive over all pairs of preferences involving W and X, W and Y, and W and Z. This can be shown by following the same pairings as above. To give one example, suppose person 1 says W P Y and person 2 says Y P W. Select the pair:

WY
XW
YX

They agree that W P X, so W SP X (PP). Also, when person 1 says X P Y and person 2 says Y P X, we have seen that X SP Y. (That W is now in the mix rather than Z does not matter because of the independence of irrelevant alternatives assumption). Therefore, W SP X and X SP Y imply W SP Y (T). Person 1 wins again.

The final step of our heuristic proof considers the case of more than two people. The key concept here is the notion of a decisive set. A subset of people is said to be decisive in the ranking of two outcomes (say, X and Y), if X SP Y when all members of the decisive set say X P Y and *everyone else* says Y P X. Once a decisive set is established, it can always be further subdivided into smaller decisive sets over other outcomes by suitably reselecting the preferences of the members inside and outside the original decisive set until the decisive set over all outcomes consists of a single person, the dictator.

The two-person example above illustrates the ability to subdivide a decisive set down to a single person. Go back to the beginning of the example when the preferences were those in the second column of six pairings and society chose Y SP Z. Think of the two lists of preferences in the second column as belonging to two subsets of the entire population, with one subset having the preferences on the left and the other on the right. Then, the subset on the left is a decisive set regarding the choice of Y and Z when the preferences over Y and Z are in the order of the second column. Perhaps society chose Y SP Z because the members on the left were in the majority.

Once the decision Y SP Z is made, then all the other possibilities in the two-person case are decided only by application of the pareto principle, transitivity, and the independence of irrelevant alternatives. Having a numerical majority, or appealing to any other criterion besides those three axioms, is irrelevant. In other words, the social decisions would hold in each subsequent example if only one person held the winning preferences and everyone else held the opposite preferences. Therefore, once a first decisive set is determined by some method such as majority voting, then some member of the decisive set is in effect a dictator. Each comparison in the examples after the first can be interpreted as person 1 having the one set of preferences and person 2 representing everyone else in the society with the opposite set of preferences on the pair under disagreement. Therefore, person 1, a member of the first decisive set, is decisive over all possibilities, a dictator. The universality axiom permits this interpretation, because the social decision process must make consistent decisions for all possible combinations of the individuals' preferences.

The implication of this form of the proof is that a consistent social decision process that generates a complete ordering of social outcomes may not result from democratic voting procedures when people disagree. It may have to be imposed by some agent who is in effect a dictator.

Cycling Preferences

Democracies are not dictators hips. Therefore, a common variation of the Arrow theorem is to assume that nondictatorship holds, along with axioms 1, 3, and 4, and then show that axiom 2 requiring a complete social ordering does not hold, in general. This variation implies that social decisions that are democratically determined do not yield a consistent set of social preferences over outcomes in general. In other words, democracies cannot expect to generate clear-cut decisions on social issues.

Consider the example of three people deciding about three different policies to divide $100 between them. The three people could be legislatures representing their constituencies. The three policies are A, B, and C, and they divide the $100 as follows:

	Person 1	Person 2	Person 3
A	$50	$20	$30
B	$30	$50	$20
C	$20	$30	$50

Suppose the three people vote according to their self-interests; they rank the policies in terms of the money they receive from each. Therefore, the individual rankings are

Person 1 : A P B P C
Person 2 : B P C P A
Person 3 : C P A P B

The social decision process is democratic: The majority rules. Unfortunately, majority voting on the three policies does not establish as best policy, even though each person has a clear set of preferences: Two of the three vote A P B (1 and 3), and two of the three vote B P C (1 and 2). Therefore, transitivity requires that A SP C, but two of the three vote C P A (2 and 3). The social preferences under majority voting are intransitive, and no clear winner can emerge when preferences are intransitive.

Often legislatures vote in pairs when there are more than three choices, with the winner of the first pairing going against the next choice. If this were done in our example, the winner would be determined by the order of the vote under majority voting:

A versus B, A wins. Then A versus C, C wins.
A versus C, C wins. Then C versus B, B wins.
B versus C, B wins. Then B versus A, A wins.

Again, no clear-cut winner emerges. The legislator who controls the order of the vote determines the winning policy.

The example illustrates a theorem about democratic voter procedures due to Duncan Black, a political scientist. Black proved that democratic voting establishes a consistent set of social preferences when people disagree if and only if the individuals' preferences are single peaked Black (1948). The problem in this example is that the preferences of person 3 are double peaked. Figure 3.6 illustrates this.

An important extension of Black's theorem considers the realistic case of voting for different options that each contain a bundle of at least two services. An example would be a vote on different local budgets that contain different proportions of expenditures on education and public safety. Black's requirement of single-peakness is almost certain to be violated in this case, implying that no consistent social consensus can emerge.

The simple example on the distributional choices gets right to the heart of the problem of determining a social

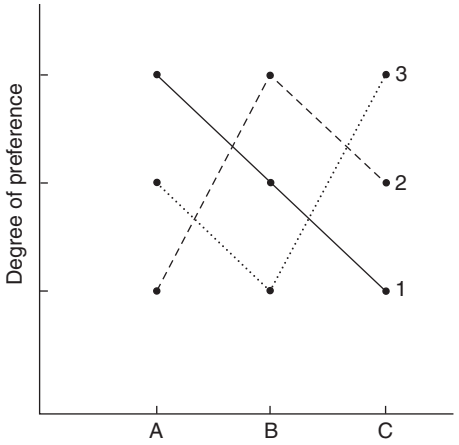

FIGURE 3.6

welfare function to resolve distributional questions in a democratic society. It suggests that no consensus social welfare function can be expected to emerge when individuals disagree about the appropriate distribution of income, as they surely do. The social welfare function that is so central to normative public sector theory may not be forthcoming in a democracy.

The Gibbard–Satterthwaite Theorem

Social decision making in democracies took a further battering in the 1970s with the publication of the Gibbard–Satterthwaite theorem. Alan Gibbard and Mark Satterthwaite proved that democratic social decisions are vulnerable to manipulation by self-serving individuals. The manipulation takes the form of lying about one's preference to achieve a more favorable social decision.[12]

Their theorem proceeds much like Arrow's. It is an exercise in cooperative game theory that focuses on the problem of choosing a single outcome based on individuals' preferences over three or more social outcomes. They posit four axioms that they believe a social decision process should possess and show that the four axioms cannot all be satisfied, in general:

1. *Universality*: The same as with Arrow; individuals can have any set of preferences over three or more outcomes.
2. *Nondegeneracy*: The social decision process cannot rule out any one outcome from being the winning outcome.
3. *Nonmanipulability*: Individuals cannot manipulate the social decision in their favor by lying about their preferences.
4. *Nondictatorship*: The same as with Arrow; the social outcome chosen cannot always be the same as the preferred outcome of one particular individual, no matter what the preferred outcomes of the other people are.

The Gibbard–Satterthwaite theorem says that these four axioms are incompatible, in general, when society is choosing among three or more outcomes. The nature of their proof can be seen by considering the two-person case above. We will not offer a complete example.

There are 36 possible pairings of the two people's preferences, and three possible social choices for each pairing (X or Y or Z), a total of 3^{36} possible choices for the social decision process to consider over the 36 pairs. Consider the first six pairings, as above:

XX	XX	XY	XY	XZ	XZ
YY	YZ	YX	YZ	YX	YY
ZZ	ZY	ZZ	ZX	ZY	ZX

We can say the following about the social choices for each pairing based on these preferences alone:

X X not Z not Z not Y?

The first two pairings generate X because X is the first choice of each. The next two pairings cannot lead to a choice of Z because it is clearly dominated by X and Y. In the third pairing, Z is the third choice of both; in the fourth pairing, Z is the second and third choice, which is lower than the combined choices for either X or Y. By the same argument, the fifth pairing cannot lead to a choice of Y. Nothing can be said about the social choice in the last pairing.

Suppose the social choice in the third pairing is X, so that society chooses in favor of person 1. If this is so, then the axiom of nonmanipulability requires that society choose X for all the pairings in the row. Person 1 is a dictator.

To see why, suppose society chooses Y for the fourth pairing, having chosen X for the third pairing. If so, then person 2 can lie about his preferences in the third pairing and represent them as Y Z X. This would make the third pairing identical to the fourth pairing, in which Y is chosen. Therefore, to prevent manipulation by person 2, society must choose X for the fourth pairing. Then, having chosen X for the fourth pairing, we now know that the last pairing cannot choose Y. If Y is not chosen in the fourth pairing when ranked second (by person 1) and first (by person 2), it will not be chosen in the last pairing when it is ranked second by both.

Next, suppose society chooses Z for the fifth or sixth pairing. If so, then person 2 can lie about his or her preferences in the fourth pairing and represent them as Z X Y or Z Y X. Society would then choose Z over X in the fourth pairing, which person 2 prefers. Therefore, to prevent

12. Gibbard (1973), Satterthwaite (1975) For more detailed discussion of the Arrow and Gibbard–Satterthwaite theorems, and of social choice generally, see Feldman (1980).

manipulation by person 2, society must choose X for the final two pairings.

In conclusion, the social choices for each of the six pairings that prevent manipulation by person 2 are

x x x x x x

Person 1 is a dictator. Conversely, preventing person 1 from being a dictator allows person 2 to manipulate the outcome.

The Gibbard−Satterthwaite theorem has three troubling implications for normative public sector theory in democratic societies.

The first is its potential devastation of the government-as-agent principle. How accurate is the information that the government collects about individuals' preferences on social issues in its role as agent? How can the government know whether people are manipulating their preferences when they vote?

The second relates to the mechanism design problem of social decision theory, which attempts to design decision-making mechanisms in which people have an incentive to reveal their true preferences. Democracies would hope that people could register their preferences voluntarily through some kind of voting mechanism. The Gibbard−Satterthwaite theorem tells us, however, that voluntary voting mechanisms can never guarantee that people will register their true preferences. Truth-revealing decision mechanisms may exist, but the theorem implies that they generally require some form of coercion by the government to implement them.

Finally, the Gibbard-Satterthwaite theorem calls into question the entire thrust of normative public sector theory. The practical value of designing truth-revealing political mechanisms may be clear enough in light of the theorem, but the normative significance is questionable. Self-interested individuals who exploit private information to their own political ends, those who cheat on their taxes and lie to government officials, fail a fundamental test of social behavior, the test of good citizenship. What is the normative significance of having the government spend time and energy designing mechanisms to prevent people from cheating and lying? In what sense is a collection of individuals a society if they are dishonest, self-serving, and manipulative and have no stake in a broader public interest? After all, many people are good citizens, honest and concerned about the public interest. They would never even think of cheating on their taxes. Should they be given higher social welfare weights (assuming they could be identified)? What is the appropriate social objective function when good citizenship is lacking in some or all? Is pareto optimality enough?

These questions underscore the main point of this section, that the politically determined social welfare function is on very shaky ground indeed in democratic societies. At the same time, the analysis of Chapter 2 indicates that mainstream public sector theory is on very shaky ground without the social welfare function.

Reactions to the Arrow and Gibbard−Satterthwaite Theorems

Public sector economists have reacted in one of three ways to the problematic nature of the social welfare function. Two are mainstream reactions; the third is associated with the public choice school.

One mainstream reaction, commonly associated with Paul Samuelson, might be termed the technocratic response: Economists should stop worrying about the social welfare function. A social welfare will emerge from the political process by whatever means; societies do make distributional judgments. Economists should simply ask the government's policy makers what the social welfare function is and then advise them how to maximize social welfare. All policy problems are constrained optimization problems consisting of objectives, alternatives, and constraints for which the given social welfare serves as the objective function. Economists can help the policy makers fill in the remaining elements of each economic policy problem, the relevant alternatives and constraints, and then describe how to solve the problem. Economists know how to solve constrained optimization problems.

A second mainstream reaction sees a more instructive role for the social welfare function. It calls for the use of flexible-form social welfare functions in normative policy exercises that allow for the full range of ethical rankings, from utilitarian to Rawlsian. The purpose of this type of analysis is to show policy makers how different ethical rankings influence optimal policy rules. This approach is not contradictory with the first approach, since the flexible social welfare function could include the government's actual social welfare function as one of the options.

Joseph Stiglitz dubbed the application of flexible form social welfare functions the "New, New Welfare Economics," because he viewed it as a direct reaction to the so-called New Welfare Economics of the 1930s and 1940s Stiglitz (1985). The older "New Welfare Economics" held that interpersonal comparisons of utility are meaningless. Economists can say nothing about situations in which some people gain and others lose, because there is no meaningful way to compare the increased utility of the gainers with the decreased utility of the losers. This older view rules out a social welfare function defined in terms of individuals' utilities and along with it any hope of an economic solution to the quest for distributive justice.

The balancing of gains and losses through redistributions is the central economic issue in achieving distributive justice. The newer breed of economists who subscribe to the "New, New Welfare Economics" want to say something about distributive justice, and in doing so

they completely reject the older view. To make the flexible-form approach operational in applied work, the researcher must specify a particular social welfare function and particular utility functions to serve as arguments in the social welfare function. Once the particular functions are specified, utility becomes cardinal and fully comparable across individuals, in direct opposition to the older view.[13]

The third reaction to the problems associated with the social welfare function, commonly associated with the public choice economists, is essentially one of indifference. Public choice economists do not care that the social welfare function is problematic because they do not accept it as a valid concept. They deny that citizens enter the political process to help resolve the public interest in distributive justice. Instead, they argue that a society's distributional policies must be understood as evolving from the desires of self-serving individuals who want to maximize their own utilities. People do not spend their political energies trying to formulate social welfare functions. There is no social welfare function, and no need for one in public sector theory.

CONCLUSION

The discussions in Chapter 3 on the distinction between first-best and second-best analyses and on the problems with the social welfare function conclude our introduction to public sector theory. The thrust of the chapter has been appropriately cautionary, a warning that the foundations underlying normative public sector theory are less firm than one would like. The chapter contains three main messages:

1. First-best analysis yields definite policy prescriptions for solving society's allocational and distributional problems, but only by adopting patently unrealistic assumptions. The main advantage of first-best analysis is the intuition it gives about the nature of the problems.
2. Second-best analysis adds a dose of realism to public sector analysis by explicitly addressing the policy and market constraints under which governments operate. But second-best analysis can never yield definitive policy prescriptions because a second-best model can incorporate only a few of the underlying constraints. Unfortunately, the results from second-best models tend to be highly sensitive to the number and form of the constraints that the analyst chooses.
3. The social welfare function is a central construct in mainstream public sector theory and the theory's only indispensable political content. It has the dual analytical tasks of resolving the question of distributive justice and selecting the one efficient allocation that maximizes social welfare from the infinity of possible efficient allocations. Yet, the social welfare function is highly problematic. Particularly troublesome are the lack of guidelines about what society's ethical judgments should be, the problem that a consistent social welfare function may not emerge under conditions that would be acceptable to a democratic society, and that democratic decision processes are susceptible to manipulation by self-serving people in the form of lying to bias outcomes in their favor. These three messages apply to all of normative economics.

A fair short summary of the state of mainstream normative public sector economics would be as follows. Virtually all mainstream normative public sector analysis relies on variations of one model, Samuelson's model of social welfare maximization. But the consensus on the underlying model has not yielded a consensus set of optimal policy prescriptions for the allocation and distribution problems that are the legitimate concerns of the government. The lack of a policy consensus stems from the inherent limitations of first-best analysis, second-best analysis, and the social welfare function discussed in Chapter 3. These limitations notwithstanding, normative public sector economics does offer important insights into all the complex allocation and distribution problems that governments have been asked to solve.

Chapters 4 through 11 in Part II turn to the first-best theory of public expenditures and taxation, in which the government is assumed to have all the necessary policy tools to reach the bliss point on the first-best utility-possibilities frontier. The first-best analysis is the core of normative public sector theory. It yields the baseline "best possible" results of public sector economics, with which the more realistic second-best results are always compared.

13. Broadway and Wildasin (1984) have an excellent discussion of restrictions on the social welfare function that make utilities comparable for policy analysis (Chapter 10, 269-277). Roemer (1996) contains a broader, deeper, and more up-to-date analysis of the same issue.

REFERENCES

Arrow, K., May 10 1973. Some ordinalist-utilitarian notes on rawls' theory of Justice. Journal of Philosophy Vol. 70 (9), 245–263.

Atkinson, A., 1973. How progressive should the income tax be? In: Parkin, M. (Ed.), Essays on Modern Economics. Longman Group Ltd., London.

Auerbach, A., Slemrod, J., June 1997. The economic effects of the tax reform act of 1986. Journal of Economic Literature Vol. 35 (2), 589–632.

Black, D., February 1948. On the rationale of group decision making. Journal of Political Economy Vol. 56 (1), 23–34.

Broadway, R., Wild sin, D., 1984. Public Sector Economics, second ed. Little, Brown and Company, Boston.

Boskin, M., May 1976. On some recent econometric research in public finance AEA Papers and Proceedings Vol. 66 (2), 102–109.

Feldman, A., 1980. Welfare Economics and Social Choice Theory. Martinus Nijhoff, Boston.

Feldstein, M., 1973. On the optimal progressivity of the income tax. Journal of Public Economics Vol. 2 (4), 357–376.

Gibbard, A., July 1973. Manipulation of voting schemes: a general result. Econometrica Vol. 41 (4), 587–601.

Gordon, S., 1980. Welfare, Justice and Freedom. Columbia University Press, New York.

Lancaster, K., Lipsey, R.G., December 1956. The general theory of the second best. Review of Economic Studies Vol. 24 (1), 11–32.

Musgrave, R., 1959. A Theory of Public Finance: A Study in Public Economy. McGraw-Hill, New York.

Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.

Roemer, J., 1996. Theories of Distributive Justice. Harvard University Press, Cambridge, MA.

Satterthwaite, M., April 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. Journal of Economic Theory Vol. 10 (2), 187–217.

Stiglitz, J., 1985. Pareto efficient and optimal taxation and the new new welfare economics. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, Vol. 2. North-Holland, New York chap. 15.

Part II

The Theory of Public Expenditures and Taxation: First-Best Analysis

Part II presents the first-best analysis of public expenditure and tax theory in the context of a market economy. Recall from the discussion in Chapter 3 that a first-best policy environment exists if the market economy is perfectly competitive and the government can use whatever policy tools that are necessary to achieve full pareto optimality and the interpersonal equity conditions of a social welfare maximum. In other words, the government can bring the economy to the bliss point on its first-best utility-possibilities frontier.

The first-best policy environment may seem unduly restrictive, but first-best analysis is the appropriate way to begin the study of the public sector. It serves as the baseline for all public sector analysis by indicating the maximum possible increase in social welfare that public policies can achieve. The social welfare implications of second-best policy prescriptions are almost always compared with their first-best counterparts. In addition, the single set of first-best assumptions permits an exploration of the essence of a technical market failure such as an externality, along with the policy required to correct it. All formal first-best analysis use variations of the general equilibrium model of social welfare maximization developed in Chapter 2, suitably modified to highlight the problem under consideration. In contrast, the restrictions added to the basic model to make the policy second best contaminate the analysis of the market failure and its solutions, with additional factors that have to do with the second-best restrictions. Finally, first-best analysis figures prominently in the history of the discipline and in much of the conventional wisdom on government policy. Virtually all public expenditure analysis before 1970 employed the first-best assumptions, as did much of the huge body of literature concerned with issues of equity in the theory of taxation. Second-best analysis in these two areas has been commonplace since then, but much of the received doctrine on public expenditures and income distribution, which appears in the current undergraduate public sector texts, comes from first-best analysis. Only the allocational theory of taxation has consistently employed second-best assumptions from the very beginning of public sector economics, simply because the welfare cost of taxation is inherently a second-best topic. As we shall discover in Part II, all interesting first-best efficiency issues relating to taxation are effectively subsumed within the optimal public expenditure decision rules.

The eight chapters in Part II are structured as follows.

Chapter 4 begins with the distribution question, one of the fundamental market failures requiring social

decisions. The chapter describes how economists use the social welfare function in applied research to determine the effects of inequality and social mobility on social welfare. Examples are drawn from the US economy.

Chapters 5—9 then turn to the two most important allocational market failures in a first-best environment: externalities and decreasing cost production. Chapters 5—8 consider the theory of externalities, with applications to US policy, and Chapter 9 presents the theory of decreasing cost production, also with US policy applications.

Chapters 10 and 11 conclude Part II with a discussion of taxes and transfers from a first-best perspective. Chapter 10 briefly reviews the first-best optimal tax and transfer rules developed to that point, stressing their limitations as guidelines for actual tax policy. The rest of the chapter is devoted to the theory of pareto-optimal redistribution, which derives normative rules for optimal redistribution without resorting to a social welfare function. pareto-optimal redistribution is the normative distribution theory favored by public choice economists, who reject the concept of a social welfare function. Chapter 11 introduces still another distributional norm, the ability-to-pay principle of taxation and transfer, which dates to Adam Smith and John Stuart Mill. The ability-to-pay principle has always been the primary guideline for tax design and tax reform in the United States and other developed market economies. The chapter begins by comparing the policy implications of the ability-to-pay principles and the interpersonal equity conditions of social welfare maximization. It then presents two applications of the ability-to-pay principle that have been featured in the public sector literature. One is how closely the US federal personal income tax adheres to the principle. The other is whether the ability-to-pay principle favors the taxation of income or consumption. The chapter concludes with two practical issues relating to the taxation of income from capital under an income tax, how to adjust for inflation, and the appropriate taxation of capital gains.

Chapter 4

The Social Welfare Function in Policy Analysis

Chapter Outline

One of the more difficult economic questions every society must face is the fundamental question of distributive justice: What is the optimal distribution of income? The question cannot be avoided. It must be answered even if the economy performs as well as it possibly can and presents no other economic problems. Moreover, the answer must come through the political process, not the market system.

Chapter 2 began the analysis of public sector economics when it presented the answer to the distribution question given by the mainstream normative public sector model in a first-best policy environment: The government should redistribute any one good or factor lump sum to satisfy the interpersonal equity conditions of an individualistic social welfare function. For some good (factor) X_k, and H individuals, redistribute such that the social marginal utilities of X are equal for all individuals, or

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hk}} = \quad h = 1, \ldots, H \tag{4.1}$$

The interpersonal equity conditions are necessary conditions for a first-best social welfare maximum, along with the pareto-optimal conditions. They are the entirety of first-best distribution theory in the mainstream model.

Chapter 4 begins Part II on first-best public sector theory with some common applications of the social welfare function in policy analysis. The applications are in the spirit of the "New, New Welfare Economics," which employs flexible-form social welfare functions to show how ethical judgments ranging from utilitarian to Rawlsian can instruct public policy. Also, because the analysis is first-best, the applications generally focus on the question of distributive justice without worrying about the inefficiencies that actual

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00004-2

redistributions of income give rise to. In other words, they assume that the pareto-optimal conditions are satisfied, unless specifically stated otherwise.

SOCIAL WELFARE AND THE DISTRIBUTION OF INCOME: THE ATKINSON FRAMEWORK

England's Anthony B. Atkinson was a pioneer of the "New, New Welfare Economics" in the early 1970s. He became interested in the possibility of making social welfare judgments based on the personal income data that England and other developed capitalist countries were collecting from surveys of the population. The US survey is the annual Current Population Survey (CPS), which began in 1947. The CPS surveys approximately 60,000 families and unrelated individuals and is the principal source of the federal government's published statistics on personal income, poverty, and other personal characteristics such as family size and education.

Atkinson's desire to meld social welfare and the income data led him to specify the social welfare function in terms of income. Write each individual's (family's) utility as a function only of income, Y_h, and the social welfare function as

$$W = W\left(U^h\left(Y_h\right)\right) \tag{4.2}$$

The relevant margin for the interpersonal equity conditions is the social marginal utility of income, $\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial Y_h}$; the product of the marginal social welfare weight, $\frac{\partial W}{\partial U^h}$; and the private marginal utility of income, $\frac{\partial U^h}{\partial Y_h}$.

The Atkinson Assumptions

Atkinson sought a very simple specification of W, one that could easily be applied to the income data and yet would capture the full range of ethical judgments from utilitarian to Rawlsian. He achieved this with three highly simplified and heroic assumptions: (1) the social welfare function is utilitarian, (2) everyone has identical tastes, and (3) utility exhibits diminishing private marginal utility of income. Atkinson's assumptions were widely adopted in applied social welfare analysis. The assumptions deserve some comment by way of justification simply because they are so strong.

Utilitarian Social Welfare

Atkinson assumed that social welfare should honor the impersonality principle, discussed in Chapter 3. The simplest way to incorporate the principle is to assume that social welfare is utilitarian, with the marginal social welfare weights always equal to one:

$$W = \sum_{h=1}^{H} U^h; \quad \frac{\partial W}{\partial U^h} = 1$$

Other researchers have chosen a less restrictive interpretation of the principle: Individuals with equal utilities should have the same social welfare weights. This variation permits flexible social welfare functions; nonetheless, it retains the strong results that follow from Atkinson's three assumptions. The impersonality principle is especially compelling in a first-best environment. Practices such as discrimination, which are used to justify affirmative action policies, do not arise in a first-best environment.

Same Preferences

This assumption is clearly false, but it can be justified in a modeling context in one of three ways. The first is to view it simply as an assumption by default. If we assume that preferences differ, how should the differences be modeled? No obvious answer comes to mind and, therefore, nothing more really need be said. Still, the assumption can be somewhat justified on other grounds.

A second possible justification is that differences in preferences should not have any influence on policy decisions (so long as tastes are not destructive). Most people would argue that policy decisions should be based on differences in peoples' circumstances rather than differences in their tastes, especially policies related to the distribution question. How much income people have is what matters, not what they choose to buy with their incomes.

A final possible justification is that people's preferences may well be quite similar if viewed from a lifetime perspective. Differences in preferences may be largely determined by different positions in the life cycle, holding circumstances constant. Single 20-year-olds have different preferences from married 50-year-olds. But the 50-year-old father may have had much the same tastes as his 20-year-old son when he was 20 years, and the 20-year-old daughter may have much the same tastes as her 50-year-old mother when she is 50 years and a mother.

Whether these last two justifications are convincing is almost beside the point. The assumption of identical tastes remains the only plausible default assumption for modeling purposes.

Diminishing Marginal Utility of Income

The assumption of diminishing marginal utility is difficult for economists to accept because diminishing marginal utility of income is neither a necessary nor a sufficient condition for any result in standard consumer theory. The best case for it is the demand for insurance under expected

utility maximization, which assumes invariance only up to linear transformations of the utility function. People who are risk averse act as if they have diminishing marginal utility of income when they pay insurance premiums to avoid exposure to risky future income streams. For example, suppose people face a 50% chance of becoming ill and losing $1000 as a result. The expected loss is $500, yet most people would be willing to pay a premium greater than $500 to insure against the possibility of the $1000 loss. This implies that their utility gain from a $500 increase in income is less than the utility loss from a $500 loss in income; the marginal utility of income is decreasing.

The insurance example refers to a single individual, however. In a social welfare context, the utility comparison is being made across individuals. The equal tastes assumption combined with diminishing marginal utility of income implies that the utility loss to the "rich" of taking $1 from them is less than the utility gain to the "poor" of giving them the $1. The notion that the marginal utility of income to the rich is less than the marginal utility of income to the poor is undoubtedly appealing to many people, especially at the extremes of income. It may even be the primary reason why the majority of people in the United States accept some redistribution of income to help the poor. But this was precisely the kind of interpersonal comparison of utility that the New Welfare Economics rejected as meaningless in the 1930s. The New, New Welfare Economics resurrected the notion of diminishing marginal utility across individuals, which had been widely accepted by political economists at the end of the nineteenth century.

The Bias Toward Equality

Atkinson's three assumptions together imply a very strong result in a first-best environment, the complete equality of incomes. The optimal policy rule from the interpersonal equity conditions is to tax and transfer lump sum to equalize the social marginal utilities of income. With the utilitarian social welfare function, however, the social marginal utilities of income are the private marginal utilities of income:

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial Y_h} = \frac{\partial U^h}{\partial Y_h} \tag{4.3}$$

with $\frac{\partial W}{\partial U^h} = 1$ for all h. Therefore, the interpersonal equity conditions imply equalizing the private marginal utilities of income:

$$\text{IE conditions (under Utilitarianism): } \frac{\partial U^h}{\partial Y_h} = \text{all } h = 1, \ldots, H$$

But assumptions two and three imply that everyone transfers income into utility by means of the same concave function. Therefore, the private marginal utilities of income are equal if and only if everyone has the same income, the mean level of income. The government should lump-sum tax everyone above the mean down to the mean, and lump-sum transfer everyone below the mean up to the mean.

Figure. 4.1 illustrates that everyone transforms income into utility according to the function $U(Y)$; the slope of $U(Y)$ is the private marginal utility of income. Suppose there are initially two classes of people, the "rich" with incomes of Y_R above the mean and the "poor" with incomes of Y_P below the mean. The $\text{MU}_{Y_R} < \text{MU}_{Y_P}$, so that aggregate utility can be increased by taxing the rich and transferring to the poor. The inequality continues to hold, and aggregate utility can be further increased, until each has reached the mean. At that point the marginal utilities are equal, the interpersonal equity conditions are satisfied, and aggregate utility is at a maximum.

Very few people would support the complete leveling of incomes, yet this result is commonplace in public sector modeling. Almost all mainstream public sector models of social welfare reach the conclusion that if redistribution is costless (i.e., lump sum), then incomes should be equalized after tax and transfer to maximize social welfare. The underlying reason for this result is that Atkinson's assumptions, heroic as they may be, have been widely accepted by public sector economists, especially the assumptions of identical preferences and diminishing marginal utility. As noted above, models do often use nonutilitarian social welfare functions with varying social welfare weights. But if the social welfare function honors the impersonality principle by giving equal social marginal welfare weights to those with equal utility levels and the other two assumptions are maintained, then the equal-incomes implication of the

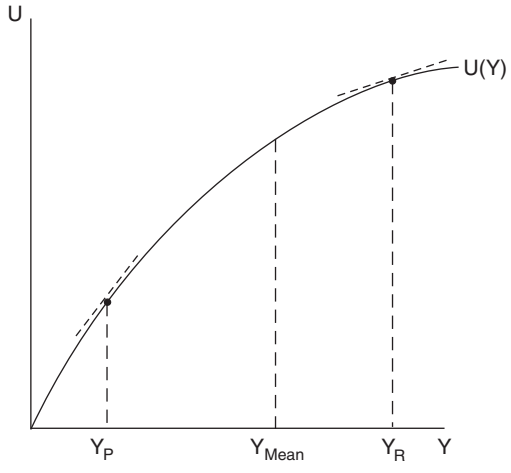

FIGURE 4.1

interpersonal equity conditions obtains with lump-sum redistributions.

In other words, Thurow's contention that there is a strong bias toward equality in the United States, noted in Chapter 1, has been incorporated into mainstream public sector theory. Thurow also maintains that the bias is so strong that inequality has to be justified. The standard justification for inequality among economists is that redistribution is not costless. Governments are forced to use distorting taxes and transfers to redistribute income, not lump-sum taxes and transfer, so that redistribution causes efficiency losses. At some point short of equality, the additional efficiency losses of further redistribution more than offset the equity gains, and redistribution should stop.

Okun's Leaky Bucket

Arthur Okun described the redistributions as if occurring with a leaky bucket, an image of the efficiency losses that has stuck in the literature Okun (1975). He imagined the rich dropping their tax dollars in a bucket, which a government official then carries to the poor. The bucket leaks, however, so that the poor receive fewer dollars than the rich had placed in the bucket. The leaks take three forms: the administrative costs to the government of taxing and transferring; the costs to the taxpayers and transfer recipients of complying with the laws, such as filing tax returns and applying for public assistance; and the deadweight efficiency losses in the marketplace as the tax and transfer programs cause buyers and sellers to face different prices for the same goods and factors.

The rich–poor example illustrates the effect of Okun's leaky bucket on the optimal amount of redistribution. With lump-sum taxes and transfers (no leaks), the redistribution continues until:

$$\mathrm{MU}_{Y_\mathrm{R}} = \mathrm{MU}_{Y_\mathrm{P}} \tag{4.4}$$

which is full equality. In terms of the effect of the redistribution on social welfare, think of the $\mathrm{MU}_{Y_\mathrm{R}}$ as the marginal cost and the $\mathrm{MU}_{Y_\mathrm{P}}$ as the marginal benefit. The redistribution continues until the marginal cost and marginal benefit are equal, the standard result. Okun's leaky bucket introduces an additional marginal cost from the three sources noted above, so that the full marginal cost of redistributing is the sum of the marginal costs borne by the rich plus the leaky bucket. Therefore, with distorting taxes and transfers (a leaky bucket), the redistribution continues until

$$\mathrm{MU}_{Y_\mathrm{R}} + \mathrm{MC}_{\mathrm{LB}} = \mathrm{MU}_{Y_\mathrm{P}} \tag{4.5}$$

where $\mathrm{MC}_{\mathrm{LB}}$ is the marginal cost of the leaky bucket. Since $\mathrm{MU}_{Y_\mathrm{R}} < \mathrm{MU}_{Y_\mathrm{P}}$ at the optimum, $Y_\mathrm{R} > Y_\mathrm{P}$. Referring again to Fig. 4.1, the initial situation pictured there could be the final equilibrium with distorting taxes and transfers. In Thurow's terms, the marginal costs of Okun's leaky bucket justify the remaining inequality.

The Atkinson Social Welfare Function

To obtain more specific results than simply equality versus inequality requires specifying a particular utility function. Atkinson chose the following utility function:

$$U^h = \frac{1}{(1-e)} Y_h^{(1-e)} \quad e = [0, \infty] \tag{4.6}$$

where e is a measure of society's aversion to inequality.[1] The utilitarian social welfare function, W, under the assumption of equal tastes for all H individuals, is

$$W = \sum_{h=1}^{H} U^h = \sum_{h=1}^{H} \frac{1}{(1-e)} Y_h^{(1-e)} \tag{4.7}$$

Atkinson chose this utility function because it is especially easy to apply in social welfare analysis. It has the following useful properties.

The Private Marginal Utilities of Income

The private marginal utilities of income, which are relevant to the interpersonal equity conditions, are simple functions of income and society's aversion to inequality:

$$\frac{\partial U^h}{\partial Y_h} = \mathrm{MU}_{Y_h} = \frac{1}{Y_h^e} \tag{4.8}$$

Marginal utility decreases with increases in Y, as required. Also, the ratio of the marginal utilities for any two people is a simple ratio of their incomes. Returning to the rich/poor example,

$$\frac{\mathrm{MU}_{Y_\mathrm{P}}}{\mathrm{MU}_{Y_\mathrm{R}}} = \left(\frac{Y_\mathrm{R}}{Y_\mathrm{P}}\right)^e \tag{4.9}$$

Therefore, the social welfare implications of any small redistribution of income are easily determined.

Society's Aversion to Inequality

Society's aversion to inequality applies directly to individual incomes in Atkinson's specification rather than to the marginal social welfare weights, which are all equal to unity. The limits of the aversion-to-inequality parameter e

1. This utility function is commonly employed in the theory of risk taking because it exhibits constant relative risk aversion, meaning that the elasticity of marginal utility with respect to income is constant. The reader can verify that the elasticity equals -e for Atkinson's utility function. See Atkinson (1983).

are the utilitarian and the Rawlsian cases. To see this, refer to the ratios of marginal utility above.

If $e = 0$, then $U^h = Y_h$ and $W = \sum_{h=1}^{H} Y_h$. Social welfare is utilitarian in income. All marginal utilities are equal to unity so that redistributing cannot raise social welfare, no matter how large the difference between Y_R and Y_P. Society is indifferent to the distribution of income.

If $e = \infty$, the ratio of marginal utilities is infinite and would be no matter what the discrepancy in income is between the rich and the poor. Because the marginal utility of the poorer of two people receives a relatively infinite weight, the poorest member of society receives an infinitely greater weight than anyone else. In effect, then,

$W = \min(Y_1, \ldots, Y_h, \ldots, Y_H)$. Social welfare is Rawlsian in incomes; society is as egalitarian as possible.

Finally, increases in e between 0 and ∞ increase the ratio of marginal utilities for any given difference in incomes Y_R and Y_P. Society's aversion to inequality increases as e increases.

Okun's Leaky Bucket Again

Atkinson's social welfare function can be applied to the CPS data on income to make social welfare inferences about the distribution of income in the United States. One of the more interesting early applications of Atkinson's framework was due to Arnold Harberger. He combined Atkinson's social welfare function with Okun's leaky bucket to argue that the United States does not care very much about inequality Harberger (1983).

At the time Harberger wrote, the average income of those in the top 10% of the income distribution was nine times greater than the average income of those in the bottom 10%. Designating the average income of those at the top Y_R and the average income of those at the bottom Y_P,

$$\frac{Y_R}{Y_P} = \frac{9}{1} \tag{4.10}$$

Suppose, said Harberger, that the aversion to inequality parameter e were equal to 1/2, fairly close to the utilitarian indifference to inequality ($e = 0$). Then,

$$\frac{MU_{Y_P}}{MU_{Y_R}} = \left(\frac{Y_R}{Y_P}\right)^{\frac{1}{2}} = \left(\frac{9}{1}\right)^{\frac{1}{2}} = \frac{3}{1} \tag{4.11}$$

With $e = 1/2$, society believes that $MU_{Y_P} = 3MU_{Y_R}$. In other words, society believes that an additional dollar of income is worth three times as much to the poorest people than to the richest people, yet it stops redistributing at a point at which the disparity in incomes between the richest and poorest is very large, nine to one. Suppose inequality is justified by the inefficiencies of redistributing as mainstream economists believe. Then these numbers imply that Okun's bucket has a huge leak, 67 cents on the dollar. Society permits a nine-to-one income disparity because it believes that only 1/3 of each additional dollar taken from the top income group in taxes would reach the bottom income group in transfers.

Harberger thought a leak of 67 cents on the dollar was absurdly large. At the time, the best estimates of the marginal dead weight loss from income taxes were on the order of 10 cents on the dollar, and everyone assumed that the administrative and compliance costs of income taxes were negligible. Therefore, Harberger concluded that the aversion-to-inequality parameter in the United States must be quite a bit less than 1/2 to justify such a large disparity in the richest and poorest incomes, that is, e is very close to zero. The United States does not care very much about inequality.

Harberger may not be correct. Estimates of the marginal costs of redistributing have been steadily increasing since he wrote. Estimates of the marginal dead weight loss from income taxes are now all over the map, but the average estimate in the literature is probably on the order of 30–40 cents on the dollar, with the high-end estimates at $1 or more. Also, economists are finding that the compliance costs of income taxes may be fairly substantial, perhaps as much as 10 cents per dollar of revenue collected. The point is that an estimate of a leak in Okun's bucket of 67 cents on the dollar, or even more, would not be considered outlandish today. Had Harberger known of these higher estimates, he might have concluded that the appropriate aversion to inequality parameter for the United States was 1/2, or even higher. At the same time, the disparity in incomes among the richest and poorest groups has also been steadily increasing; it now exceeds 15 to 1. The increasing inequality would tend to lower the estimate of e for the United States. Whatever the true value for e may be, Harberger's calculations illustrate that the Atkinson social welfare framework offers a very convenient first-pass means of thinking about the equity-efficiency trade-off in redistributing income.[2]

Social Welfare Indexes of Inequality

Atkinson was particularly interested in the social welfare implications of inequality. His approach was to incorporate the social parameter e, the aversion to inequality, directly into an index of income inequality. He was widely followed in this and spawned a huge literature on inequality measurement.

The most popular way of presenting data on the distribution of income has always been the Lorenz curve and its associated Gini coefficient. The Lorenz curve compares

2. A more recent test of students to determine their aversion to inequality found that it was very low, around 0.25, much as Harberger had surmised. See Amiel et al. (1999).

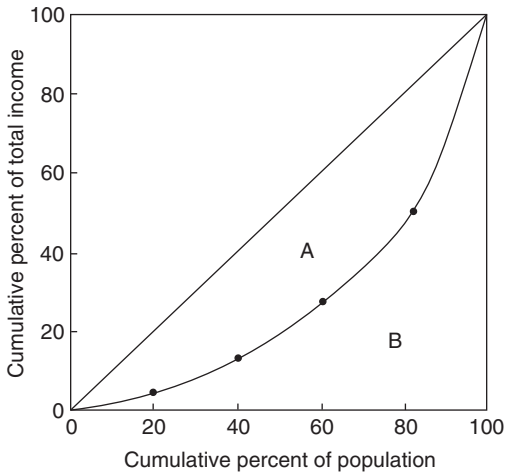

FIGURE 4.2

the cumulative percent of the total income with the cumulative percent of the total population, when individuals (or families) are ordered from lowest to highest income (Fig. 4.2). The Lorenz curve is typically drawn inside a square. The bottom of the square, the horizontal axis, records the cumulative percent of the total population; the sides of the square, the vertical axes, record the cumulative percent of the total income earned by each cumulative percent of the population.

Every Lorenz curve must begin in the lower left-hand corner (0% of the population earns 0% of the total income) and end at the top right-hand corner (100% of the population earns 100% of the total income). The diagonal of the square is the line of perfect equality; each x% of the population earns x% of the total income. Actual Lorenz curves lie below the diagonal because incomes are unequally distributed. The further below the diagonal the Lorenz curve lies, the more unequal the distribution of income.

The distribution on income by quintiles for US families and households in 2011 is presented in Table 4.1, and the Lorenz curve in Fig. 4.2 represents the household data.

The Gini coefficient is the ratio of the area A between the Lorenz curve and the diagonal to the entire area under the diagonal, $A + B$:

$$\text{Gini} = \frac{A}{(A+B)} \tag{4.12}$$

Its values lie between 0 ($A = 0$, perfect equality) and 1 ($B = 0$, perfect inequality in the sense that one person has all the income).

Atkinson used his social welfare framework to think about the following problem. Consider two different distributions of income that have the same means:

$$Y^A = \left(Y_1^A, \ldots, Y_h^A, \ldots, Y_H^A\right) \text{ and}$$
$$Y^B = \left(Y_1^B, \ldots, Y_h^B, \ldots, Y_H^B\right)$$

Let W^A be the social welfare associated with Y^A, and W^B be the social welfare associated with distribution Y^B. Assume, as above, identical tastes and diminishing private marginal utility of income (that is, social welfare cannot be utilitarian in income). Can anything of a general nature be said about W^A versus W^B?

The answer is yes, under certain conditions. Atkinson proved, in the case of equal means, that $W^B > W^A$ for all values of $e \neq 0$ if and only if the Lorenz curve for distribution Y^B lies everywhere inside the Lorenz curve for distribution Y^A Atkinson (1970). Y^B is said to Lorenz dominate Y^A. The intuition is that under Lorenz dominance the more equal distribution can be obtained from the less equal distribution by a top-down redistribution from those with higher income to those with lower income. Such a top-down redistribution must increase individualistic social welfare under diminishing marginal utility of income, because the utility gains of those with lower income exceed the utility losses of those with higher incomes per dollar of income transferred. Atkinson's theorem was the first direct link between social welfare and the Lorenz curve representation of inequality.

Atkinson's theorem has limited applicability for two reasons. One is that two distributions often have different mean incomes and the other is that the two Lorenz curves may cross.

TABLE 4.1 Personal Distribution of Income in the United States: Families and Households, 2011

	Quintile				
	Bottom	Second	Third	Fourth	Top
	20%	20%	20%	20%	20%
Percentage of total income for families	3.8	9.3	15.1	23.0	48.9
Percentage of total income for households	3.2	8.4	14.3	23.0	51.1

U.S. Census Bureau (2013). Available on the Bureau's Web site.

Generalized Lorenz Dominance

Tony Shorrocks extended Atkinson's theorem to distributions with different mean incomes by defining a "generalized" mean-augmented Lorenz curve of the following form Shorrocks (1983). Represent the standard Lorenz curve for the $100p\%$ poorest individuals as

$$L(p) = [Y(1) + \ldots + Y(j)]/H\mu \qquad (4.13)$$

where H is the total population; $j = 1,\ldots, H$; $p = j/H$; and μ is the mean income. Shorrocks' generalized mean-augmented Lorenz curve is

$$GL(p) = \mu L(P) = [Y(1) + \ldots + Y(j)]/H \qquad (4.14)$$

Points on $GL(p)$ are a hybrid per capita income measure in which the numerator is the sum of the incomes of the $100p\%$ poorest individuals and the denominator is the total population. Consequently, the vertical axis of the generalized Lorenz curve runs from 0 to the mean level of income, as pictured in Fig. 4.3. Also, the diagonal is still the line of perfect equality. If income were equally distributed, any $100p\%$ of the population would have a hybrid per capita income equal to $100p\%$ of the mean income.

Shorrocks showed that for two income distributions, Y^A and Y^B, $W^B > W^A$ for all $e \neq 0$ if and only if $GL_B(p) > GL_A(p)$ for all p [0, 1]. That is, the generalized Lorenz curve for Y^B lies everywhere above (Lorenz dominates) the generalized Lorenz curve for Y^A. Notice that one requirement for Y^B to have higher social welfare is that it must have a larger mean: $\mu^B > \mu^A$ at $p = 1$. The intuition behind Shorrocks' theorem, then, is that Y^B has higher social welfare because $100p\%$ of the poorest people always have a higher share of a larger mean income relative to Y^A.

Crossing Lorenz Curves

Unfortunately, Lorenz and generalized Lorenz curves may cross, in which case, Atkinson's and Shorrock's theorems do not apply. Distributions whose Lorenz curves cross require a specific social welfare function to determine which has higher social welfare because different values of e ($\neq 0$) can generate different social welfare rankings. Fig. 4.4 illustrates for the standard Lorenz curve. The Lorenz curves in the figure cross once, at 15% of the total population. A social welfare function that gives a large weight to the bottom 14% of the population might prefer distribution Y^A, because the bottom 14% receive a higher share of the total income under Y^A. Conversely, a social welfare function that gives a large weight to the bottom 16% of the population might prefer distribution Y^B, because the bottom 16% receive a higher share of the total income under Y^B. Society's aversion to inequality matters in ranking the two distributions.

Atkinson's Index of Inequality

Atkinson proposed to rank instances of crossing Lorenz curves by constructing an index of inequality that directly incorporates society's aversion to inequality into the index. As noted earlier, his proposal stimulated a large number of imitators. Social welfare-based indexes of inequality have been widely used for determining the incidence of government expenditures and taxation from a social perspective, as well as in the analysis of income distributions. We will discuss some of the incidence applications in later chapters.

Atkinson's index of inequality follows directly from his social welfare framework and is constructed as follows. Using Eqn (4.7), calculate the level of social welfare, W^A,

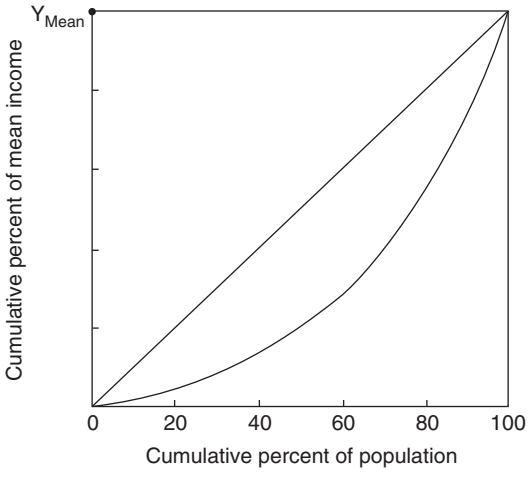

FIGURE 4.3

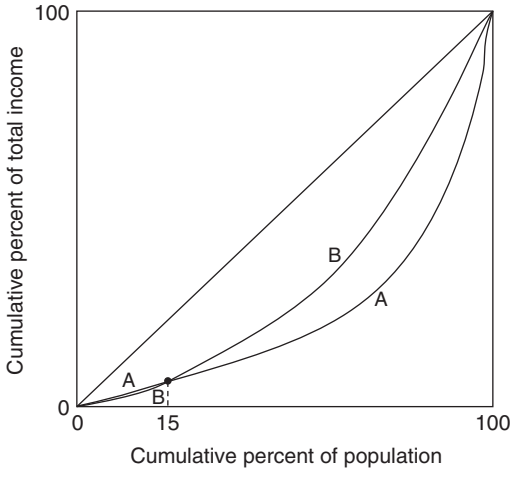

FIGURE 4.4

implied by the distribution Y^A, for a given value of e. Next, determine the amount of income that, if given equally to everyone, would generate the same level of social welfare, W^A, as the given distribution. Atkinson called this income the "equally distributed equivalent" income, labeled Y_{ede}. For Atkinson's social welfare function, Y_{ede} is a solution to the equation:

$$W^A = \frac{H}{(1-e)} Y_{\text{ede}}^{(1-e)} \tag{4.15}$$

Finally, use Y_{ede} to form the following index of inequality:

$$I(e) = 1 - \left[\frac{Y_{\text{ede}}}{Y_{\text{mean}}}\right] \quad I(e): \ [0,1] \tag{4.16}$$

Note that the index depends on the aversion to inequality parameter e.

$I(e)$ has two attractive properties. First, $I = 0$ represents "perfect equality" and $I = 1$, "perfect inequality," in line with most indexes of inequality. $I = 0$ either if everyone has the same income so that $Y_{\text{ede}} = Y_{\text{mean}}$, or if $e = 0$ and society is indifferent to inequality because social welfare is utilitarian in income. $I = 1$ if social welfare is Rawlsian, with $e = \infty$. Second, $I(e)$ has a natural interpretation of the social cost of inequality for given values of e. Suppose $I(e) = 0.25$. Then $\left[\frac{Y_{\text{ede}}}{Y_{\text{mean}}}\right] = 0.75$. In other words, the index says that society could have the same level of social welfare with only 75% of the total income if income were equally distributed. Twenty-five percent of the total income can be viewed as the social cost of the given inequality. Note, finally, that any two distributions can be ranked using Atkinson's index of inequality.[3]

Inequality versus Social Welfare: Sen's Critique

Amartya Sen, the 1998 Nobel Laureate in Economics, is one of the leading economic theorists working on the problems of inequality, poverty, and social justice. He has been highly critical of all attempts to incorporate social welfare into indexes of inequality. Sen argues that social welfare and inequality are both primitive concepts, meaning that one cannot be derived from the other as Atkinson and others have tried to do. This is most easily seen if social welfare is utilitarian in terms of income ($e = 0$). Suppose society consists of two people and consider the two distributions:

$$Y^A = (\$5, \$5) \text{ and } Y^B = (0, \$10).$$

Everyone would say that Y^B is the more unequal distribution, yet they both yield the same social welfare with a utilitarian social welfare function.

Sen points out that the fundamental inconsistency between social welfare and inequality is not limited to the knife-edge utilitarian case. It is a more general problem. To see this, suppose that $e > 0$ and the incomes of the two people are unequal. Consider a reverse Robin Hood transfer of \$1 from the poorer person to the richer person. Ask what happens as e decreases to the inequality of income, the inequality of utility, and the change in social welfare. Use Atkinson's social welfare framework to make the comparisons.

Inequality of Income

The inequality of income does not change. All straight measures of income inequality, such as the Gini coefficient, do not incorporate social welfare and are therefore independent of e.

Inequality of Utility

The inequality of utility increases. The change in the inequality of utility from the transfer is the sum of marginal utilities of income, $\text{MU}_{Y_R} + \text{MU}_{Y_P}$. The utilities of the two people are being driven further apart by the transfer. But $\text{MU}_Y = (1/Y^e)$, which increases as e decreases for all Y. Therefore, the sum of the marginal utilities increases as e decreases.

Social Welfare

The change in social welfare may decrease. The change in social welfare from the transfer is the difference in the marginal utilities of income, $\text{MU}_{Y_R} - \text{MU}_{Y_P}$. Giving the higher income person one more dollar increases social welfare by MU_{Y_R}; taking the dollar from the poor decreases social welfare by MU_{Y_P}. The difference in marginal utilities, $(1/Y_R^e) - (1/Y_P^e)$, can become less negative as e decreases for certain ranges of incomes and e, as the reader can verify.

Sen's examples suggest that attempts to infer changes in social welfare from changes in inequality are problematic, the more so if a nation cannot reach a consensus on its aversion to inequality Sen (1982).

The Atkinson Framework and Inequality in the United States

John Bishop, John Formby, and James Smith (BFS) applied the Atkinson framework to the CPS income data from 1967 to 1986 to track changes in social welfare over those 20 years. They found that Lorenz curves calculated from the CPS data crossed in 7 of the 20 years. The CPS data are

3. Peter Lambert has written an excellent survey of the relationship between income measures of inequality and social welfare Lambert (1993).

just a sample of the entire US population, however. Given the sample variance of incomes, BFS developed a test of statistical significance for Lorenz curve crossing. They concluded that the Lorenz curves crossed only once, from 1973 to 1974, on the basis of statistical significance. In all other year-to-year comparisons, the new Lorenz curve was either entirely inside or entirely outside the old Lorenz curve in a statistical sense.

The change in social welfare from one year to the next depends on the change in the mean level of income and the change in inequality as measured by the year-to-year positions of the Lorenz curves. An increase in the mean increases social welfare and an increase in inequality decreases social welfare (and vice versa). BFS discovered three distinct periods in the data, each with a consistent pattern in the year-to-year changes:

1967–78—Social welfare increased; the mean increased and inequality decreased.
1979–83—Social welfare decreased; the mean was essentially constant and inequality increased (the increase in inequality was "relatively massive").
1983–86—Social welfare increased; the mean increased, and inequality was essentially constant.

Their findings produced one major surprise. The deep recession of 1974/75 did not prevent a continuing decrease in inequality that had been ongoing for 7 years, whereas the deeper recession of 1981/82 led to a massive increase in inequality. Why inequality responded so differently to the two recessions remains an intriguing open question Bishop et al. (1991).

SOCIAL WELFARE AND CONSUMPTION: THE JORGENSON ANALYSIS

Dale Jorgenson provided an important extension of Atkinson's social welfare analysis shortly after Atkinson's work appeared. He developed a method for linking measures of social welfare to people's consumption patterns rather than their incomes Jorgenson (1990).

Econometric demand analysis of aggregate consumption data was well established by the mid-1970s, including estimation of the aggregate consumption function and major categories such as food, clothing, and transportation. Panel data sets that permit more microeconometric demand analysis were not yet available. Jorgenson's idea was to meld econometric demand analysis with the social welfare function by using the estimated demand equations for the major consumption categories to track changes in social welfare over time. His approach has three distinct steps:

1. Posit individual utility functions defined over a set of consumer goods and derive demand equations for the goods from the utility functions.
2. Estimate the demand equations in a manner that allows for recovery of the unknown parameters of the utility functions.
3. Use the estimated utility functions as the arguments of a flexible-form social welfare function that registers society's aversion to inequality, and track changes in social welfare over time.

The Estimating Share Equations

Jorgenson begins by assuming that each household, h, has an indirect utility function, V^h, defined over three sets of arguments: the prices of the various consumer goods, $\overrightarrow{P}_k$; the household's income, M^h; and a vector of household characteristics, $\overrightarrow{A}^h$, such as family size, age of the head of household, and where the household resides:

$$V^h = V^h\left(\overrightarrow{P}_k; M^h; \overrightarrow{A}^h\right) \quad h = 1, \ldots, H \tag{4.17}$$

The parameters of V^h are assumed to be equal for all households and constant over time. That is, households have identical, unchanging tastes. They also face the same vector of consumer prices. Therefore, the differences in households' utilities are due entirely to differences in their circumstances, that is, their incomes and characteristics.

Jorgenson employed the transcendental logarithmic (translog) indirect utility function to approximate the true indirect utility function. The translog is a second-order Taylor series expansion in the logs of the independent variables around their means (each independent variable is scaled by dividing by its own mean, so that the log at each variable's mean is zero). For example, the translog approximation of V^h assuming N prices and a single characteristic A^h, is

$$\begin{aligned}\ln V^h = {} & \sum_{i=1}^{N} \alpha_i \ln P_i + \alpha_M \ln M^h + \alpha_A \ln A^h \\ & + 1/2 \sum_{i=1}^{N} \sum_{j=1}^{N} \beta_{ij} \ln P_i \ln P_j + \sum_{i=1}^{N} \beta_{iM} \ln P_i \ln M^h \\ & + \sum_{i=1}^{N} \beta_{iA} \ln P_i \ln A^h + 1/2\beta_{MA} \ln M^h \ln A^h \\ & + 1/2\beta_{MM}\left(\ln M^h\right)^2 + 1/2\beta_{AA}\left(\ln A^h\right)^2\end{aligned} \tag{4.18}$$

The estimating equations are obtained by taking log derivatives of the translog function with respect to each of the prices and income:

$$\frac{\partial \ln V^h}{\partial \ln P_k} = \frac{\partial V^h}{\partial P_k}\frac{P_k}{V^h} \quad k = 1, \ldots N \tag{4.19}$$

From Roy's identity, $(\partial V^h/\partial P_k) = \lambda^h X_{\text{hk}}$, where λ^h is the marginal utility of income for household h, and X_{hk} is the consumption of good k by household h. Therefore,

$$\frac{\partial \ln V^h}{\partial \ln P_k} = \lambda^h \frac{P_k X_{\text{hk}}}{V^h} \quad k = 1, \ldots, N \tag{4.20}$$

Similarly,

$$\frac{\partial \ln V^h}{\partial \ln M^h} = \frac{\partial V^h}{\partial M^h}\frac{M^h}{V_h} = \frac{\lambda^h M^h}{V^h} \tag{4.21}$$

Dividing Eqns (4.20) by (4.21) yields:

$$\frac{\left(\frac{\partial \ln V^h}{\partial \ln P_k}\right)}{\left(\frac{\partial \ln V^h}{\partial \ln M^h}\right)} = \frac{P_k X_{\text{hk}}}{M^h} \quad k = 1, \ldots, N \tag{4.22}$$

the expenditure share of good k for household h. The expenditure shares become the dependent variables in the demand estimation. The advantage of using the expenditure shares is that the researcher does not have to worry about separating out prices from quantities.

Next, write out the price and income derivatives of the translog function to see the full system of estimating equations:

$$\frac{\partial \ln V^h}{\partial \ln P_k} = \alpha_k + \sum_{i=1}^{N} \beta_{ik} \ln P_i + \beta_{kM} \ln M^h + \beta_{kA} \ln A^h \quad k = 1, \ldots, N \tag{4.23}$$

$$\frac{\partial \ln V^h}{\partial \ln P_k} = \alpha_M + \sum_{i=1}^{N} \beta_{iM} \ln P_i + \beta_{MM} \ln M^h + \beta_{MA} \ln A^h \tag{4.24}$$

Dividing each of the N Eqns (4.23) by (4.24) yields the entire system of share equations to be estimated. The left-hand sides (LHSs) are the N expenditure shares. The right-hand side (RHS) is a nonlinear combination of the independent variables and the coefficients of the translog utility function. The system can be estimated by nonlinear estimating techniques if the data on the individual households are available.

The required microdata were not available to Jorgenson, however. He had individual household data on income and characteristics from surveys such as the annual CPS, but he only had aggregate US data on the expenditure shares for most years. The question, then, was whether the parameters of the individual household's translog utility function could be recovered from an estimation on the aggregate expenditures shares. The answer in general is no, without further restrictions on the utility parameters beyond the restrictions implied by utility maximization.

The problem without further restrictions can be seen as follows. Think of each coefficient in the share equation as a coefficient in the price derivative Eqn (4.23) divided by the entire RHS of the income derivative Eqn (4.24). The share coefficients defined in this way are functions of all the prices, income M^h, and the single individual characteristic A^h. Next, compute the aggregate shares from the individual shares. The aggregate shares are weighted averages of the individual shares, with the weights equal to each household's share of total income:

$$\sum_h \frac{M^h W_{hk}}{\sum_h M^h} = \sum_h \frac{M^h P_k X_{hk}}{\sum_h M^h} = \frac{\sum_h P_k X_{hk}}{\sum_h M^h} = W_k^{Agg} \tag{4.25}$$

The aggregate share coefficients as defined above would vary depending on the distribution of income and the characteristic across households. They would not be the same as the coefficients from each household's share equation, and they must be the same to recover the individual utility parameters in the estimation.

The weakest restriction that makes the aggregate and individual share coefficients the same is that the individual expenditure shares are linear functions of the household's income and characteristic. This in turn requires that the RHS of Eqn (4.24) be independent of a household's income and characteristic, or that $\beta_{MM} = \beta_{MA} = 0$ for the system as written above. Jorgenson refers to these two restrictions as the exact aggregation restrictions. With these two restrictions, the individual share coefficients defined by dividing Eqn (4.23) by Eqn (4.24) as above are functions only of the prices. Write:

$$W_{hk} = \alpha'_k + \sum_i \beta'_{ik} \ln P_i + \beta'_{kM} \ln M^h + \beta'_{kA} \ln A^h \quad k = 1, \ldots, N; \; h = 1, \ldots, H \tag{4.26}$$

where the α', β' coefficients are the α, β coefficients in Eqn (4.23) divided by Eqn (4.24) with the aggregate aggregation restrictions imposed.

The aggregate shares are

$$\begin{aligned} W_k^{Agg} &= \frac{\sum_h M^h W_{hk}}{\sum_h M^h} \\ &= \alpha'_k + \sum_i \beta'_{ik} \ln P_i + \beta'_{kM} \frac{\sum_h M^h \ln M^h}{\sum_h M^h} \\ &\quad + \beta'_{kA} \frac{\sum_h M^h \ln A^h}{\sum_h M^h} \quad k \\ &= 1, \ldots, N \end{aligned} \tag{4.27}$$

The only difference in the individual and aggregate share equations is the independent variables. The aggregate shares are regressed on income-weighted shares of individual household's income and characteristic, so that the aggregate shares depend on the joint distribution of incomes and characteristics. But, the coefficients in the aggregate and individual share equations are the same. Therefore, the parameters of the individual translog utility function can be recovered from estimates of the aggregate share equations, as required.

The remaining issue is to ensure that the estimated system of share equations, Eqn (4.27), is consistent with consumer theory, so that the system can be derived from a translog indirect utility function of the form of Eqn (4.18). For this to be true, the coefficient estimates must satisfy the integrability conditions on demand functions, which requires imposing a large number of a priori restrictions on the coefficients both within and across equations. To give one example, the matrix of the price coefficients B_{pp} must be symmetric. Jorgenson shows that the integrability conditions, combined with the exact aggregation restrictions, lead to a translog utility function of the form (in vector notation)

$$\ln V^h = \ln p'\alpha_p + 1/2 \ln p' B_{pp} \ln p - D(P) \ln \left[M^h / m^0(P, A^h)\right] \tag{4.28}$$

where $D(p)$ is the denominator of the share equations, Eqn (4.24), with the exact aggregation restrictions imposed, and the normalization $\alpha_p' 1 = -1$. $m^0(P, A^h)$ is a translog household equivalence scale that captures the effect of the household's characteristics on its utility level. It can be interpreted as the number of household equivalent members, so that the bracketed expression at the end of Eqn (4.28) is the per capita expenditure defined in terms of household equivalent members. Equation (4.28) is the central equation used to track changes in social welfare over time.[4]

Social Welfare

Once the translog utility parameters have been estimated, each household's indirect utility is determined by substituting the values of the prices, the household's income, and the household's characteristic(s) in Eqn (4.18). Social welfare is then a function of the households' indirect utility functions, $\ln V^h$.

Jorgenson assumed that social welfare should depend positively on the mean level of utility and negatively on two factors: the inequality of households' utilities around the mean and Atkinson's aversion to the inequality parameter e. He chose a social welfare function of the general form

$$W = \overline{\ln V} - g\left(\overline{\ln V} - \ln V^h; e\right) \tag{4.29}$$

$\overline{\ln V}$ is a weighted average of the logs of the indirect utilities, with the weights equal to the household equivalence scale, $m^0(P, A^h)$:

$$\overline{\ln V} = \frac{\sum_h m^0(P, A^h) \ln V^h}{\sum_h m^0(P, A^h)} \tag{4.30}$$

As in Atkinson, $e = [0, \infty]$, with $e = 0$ representing the utilitarian case of no concern for inequality and ∞ representing the most egalitarian Rawlsian case. $g(\ )$ is a complex function with the following properties:

1. $g1$, $g_2 > 0$; g increases and social welfare decreases if either inequality increases or society's concern for inequality increases.
2. $g = 0$ if either $V^h = V$ for all h (there is no inequality) or $e = 0$ (society is unconcerned about inequality). Notice that, under either condition, W is maximized and equal to $\overline{\ln V}$ for a given sum of the households indirect utilities.
3. g yields equal-weighted social marginal utilities and satisfies the impersonality principle, in the sense that two people with the same level of indirect utility have the same effect on g and, therefore, on W.

Given W, the researcher can track social welfare over time as a function of prices, P_t; households' incomes, M_t^h; households' characteristics, A_t^h; and society's aversion to inequality, e_t, which might also change over time. The only maintained hypotheses are that individual preferences remain constant over time (the estimated coefficients of the indirect utility function are unchanged) and that the form of W also remains the same.[5]

4. The discussion in the text ignores a number of other econometric issues associated with estimating the system of share equations so that the system is consistent with consumer theory, such as the nature of the error-covariance matrix for the entire system, and further coefficient restrictions that Jorgenson imposes to reduce the number of coefficients to be estimated or to allow him to ignore parameters in $\ln V^h$ in Eqn (4.18) that do not appear in the system of share Eqns (4.26) and (4.27). Our goal is to give an overview of Jorgenson's approach without getting bogged down in the econometric details. A complete discussion of the estimation of the translog indirect utility function (4.18) can be found in Jorgenson (1990). Another excellent and readily accessible overview of the Jorgenson approach to measuring social welfare is contained in Jorgenson (1985). See also the cautionary notes by Fisher, Blackorby, and Donaldson on the implicit assumptions behind Jorgenson's use of household equivalence scales and the cardinalization of utility when making interpersonal comparisons (Fisher, 1987) (Blackorby and Donaldson, 1988). A good general reference on social welfare measurement is Section 3 of Slesnick (1998).

5. The social welfare rankings over time implied by W are invariant to linear transformations of the indirect utility functions of the form $V^{h'} = a + bV^h$, with a and b the same for all households. The indirect utility functions are cardinal and fully comparable under this condition. The exact form of W, along with a complete discussion of its properties, is in Jorgenson (1990).

Income Measures of Social Gain and Loss

Jorgenson's final contribution was to propose income measures of gains and losses in social welfare comparable to the Hicksian compensating and equivalent variations (HCV and HEV, respectively) that are used to measure gains and loses of individual well-being. The Hicksian measures are derived from the consumer's expenditure function. Jorgenson derives his income measures from a concept that he calls the social expenditure function. A brief review of the consumer's expenditure function will be useful to understand Jorgenson's analogous social expenditure function and his income measures of social welfare gains and losses.

The Expenditure Function, HCV, and HEV[6]

The expenditure function follows directly from the dual to the standard consumer problem of maximizing utility subject to a budget constraint. The dual problem is to minimize "expenditures" subject to utility being held constant:

$$\min_{(X_i)} \sum_{i=1}^{N} q_i X_i$$
$$\text{s.t. } U(X_i) = \overline{U}$$

where the X_i are the quantities and the q_i are the prices of the goods and variable factors. "Expenditures" is understood to mean expenditures on all goods and services less income from all variable factors, given that the X_i include all variable factors supplied. The first-order conditions of the dual solve for goods demand and factor supply curves of the form:

$$X_i^c = X_i(\vec{q}; \overline{U}), \quad \text{for } i = 1, \ldots, N \tag{4.31}$$

These are compensated demand and supply curves; they show they consumer's response to price changes given that utility is held constant at $\overline{U}$. (By contrast, the ordinary market demand curves show responses to price changes given that lump-sum income is held constant.)

To form the expenditure function, replace the X_i in the objective function of the dual with the compensated supply and demand relationships (4.31) to obtain:

$$M(\vec{q}; \overline{U}) = \sum_{i=1}^{N} q_i X_i^c(\vec{q}; \overline{U}) \tag{4.32}$$

The function, M, is the consumer's expenditure function, defined solely in terms of prices and a constant utility level. Since the function is derived from the dual of the standard consumer problem, it is certainly a valid representation of consumer's preferences. Furthermore, the relationship between a primal problem and its dual guarantees that the value of the expenditure function equals the consumer's lump-sum income when $\overline{U}$ is set at the maximum utility level obtained from solving the standard utility maximization problem.

Hicks' Compensating and Equivalent Variations

Economists are naturally interested in knowing whether changes in prices increase or decrease the consumer's utility and by how much. Direct utility measures are not useful for this purpose, however, because they require cardinality, the choice of a particular utility index, even though consumer's demands (and factor supplies) are invariant to monotonic transformations of the utility index. Rather, one wants an income measure of gains and losses that is invariant to monotonic transformations of the utility index. The proper income measure is based on the notion of compensation, or indifference: How much lump-sum income (payment) is required to keep the consumer indifferent to the change in prices? The expenditure function provides the basis for this measure, because for any price vector $\vec{q}$, $M(\vec{q}; \vec{U})$ gives the minimum expenditures, or lump-sum income, required to keep the consumer at an arbitrarily selected utility level, $\overline{U}$.

To relate the expenditure function to the standard treatment of income-compensation criteria in terms of consumer indifference curves, consider a two-good example in which all factor income is lump sum because of fixed factor supplies. Suppose that the consumer is originally in equilibrium at point A on I_1 in Fig. 4.5, with relative prices q_1/q_2 indicated by the slope of the budget line tangent to I_1 at A. Suppose the price of X_2 increases, resulting in a new equilibrium at point B in Fig. 4.5 (assume the consumer's lump-sum income remains unchanged).

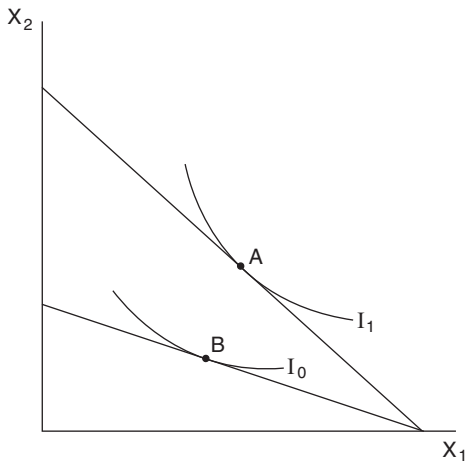

FIGURE 4.5

6. This section can be skipped by students familiar with the expenditure function and the Hicksian compensation measures of individual welfare.

The parallel distance between I_1 and I_0 gives an income measure of the welfare loss caused by this price increase. The distance is invariant to monotonic transformations of the utility index because the indifference curves are invariant to these transformations. In general, there is an infinity of possible income measures since the parallel distance between I_1 and I_0 varies depending on the slope of the parallel lines used to measure the distance. The two most popular, and natural, choices used to measure the parallel distance are the slopes corresponding to the initial and final price vectors.

In Fig. 4.6, the parallel distance from point B to C gives the additional lump-sum income necessary to compensate the consumer for the new set of prices. With this additional income the consumer would remain on I_1 (at point C) despite the higher prices and would therefore be indifferent to the new prices. This income measure is HCV. The parallel distance from point A to D gives the lump-sum income the consumer would be willing to sacrifice to maintain the old set of prices. By giving up this income, the consumer would remain on I_0 (at point D) despite facing the original prices and would therefore be indifferent to returning to the original prices. This income measure is HEV.

In general, the value of the expenditure function at the new price vector and the original utility level, $M(\overrightarrow{q}_1; \overline{U}^0)$, measures the lump-sum income necessary for indifference to the new price vector. Subtracting off the consumer's actual amount of lump-sum income, I^0 (assumed unchanged in this example), gives the HCV [7]:

$$\begin{aligned} \text{HCV} &= M\left(\overrightarrow{q}_1; \overline{U}^0\right) - I^0 \\ &= M\left(\overrightarrow{q}_1; \overline{U}^0\right) - M\left(\overrightarrow{q}_1; \overline{U}^1\right) \end{aligned} \tag{4.33}$$

The expenditure function defined at the original set of prices and the new utility level, $M(\overrightarrow{q}_0, \overline{U}^1)$, measures the lump-sum income necessary for indifference to the new utility level but at the original price vector. Subtracting off the consumer's actual lump-sum income gives the HEV:

$$\text{HEV} = M\left(\overrightarrow{q}_0, \overline{U}^1\right) - I^0 = M\left(\overrightarrow{q}_0, \overline{U}^1\right) - M\left(\overrightarrow{q}_0, \overline{U}^0\right) \tag{4.34}$$

The HEV is the preferred measure when comparing three or more situations because it is always calculated using the original price vector. By standardizing on the prices, it gives an unambiguous welfare ordering of the situations. The HEV is said to be a money metric of utility because of this ordering property. In contrast, the HCV makes pairwise comparisons of the situations using different price vectors each time. The pairwise comparisons at the different prices may not yield a transitive ordering of the utilities.[8]

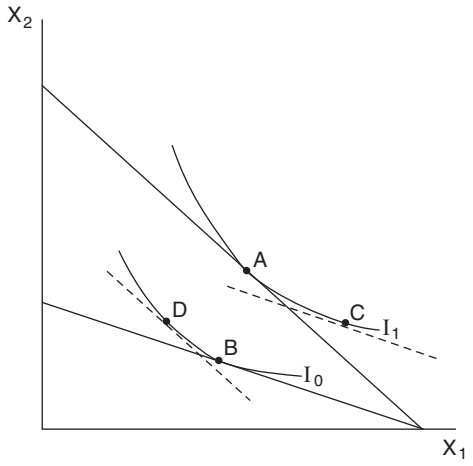

FIGURE 4.6

7. As defined in Eqn (4.33), the HCV is a loss measure. If, as in this example, goods prices should rise, then the income necessary to compensate the consumer will generally exceed the income actually available, and the HCV as written will be positive. Since the consumer is surely worse off, this positive value gives an income measure of his welfare loss. A welfare gain would be measured negatively. Some writers reverse the signs so that a gain is measured positively.

8. If both price and lump-sum income change simultaneously, these two HCV and HEV expressions have to be modified as follows. HCV becomes

$$\text{HCV} = M\left(\overrightarrow{q}_1; \overrightarrow{U}^0\right) - I^1 \tag{4.33a}$$

The HCV is the lump-sum income necessary to keep the consumer at the original utility level, given the new price vector, less the lump-sum income actually available at the new level, I^1. Alternatively, since $M(\overrightarrow{q}_0; \overline{U}^0) = I^0$, from the duality of the consumer problem,

$$\text{HCV} = \left(I^0 - I^1\right) - \left[M\left(\overrightarrow{q}_0; \overline{U}^0\right) - M\left(\overrightarrow{q}_1; \overline{U}^0\right)\right] \tag{4.33b}$$

In terms of changes in lump-sum income, then, the consumer's gain or loss is the actual change in lump-sum income less the additional income required to keep him or her indifferent to the price changes, measured at the original utility level.

Similarly, the HEV is now the income required to keep the consumer at the new utility level with the old price vector, less the income actually available in the initial situation:

$$\text{HEV} = M\left(\overrightarrow{q}_0; \overline{U}^1\right) - I^0 \tag{4.34a}$$

Alternatively, since $M(\overrightarrow{q}_1; \overline{U}^1) = I^1$ from duality,

$$\text{HEV} = \left(I^1 - I^0\right) - \left[M\left(\overrightarrow{q}_1; \overline{U}^1\right) - M\left(\overrightarrow{q}_0; \overline{U}^1\right)\right] \tag{4.34b}$$

The HEV is the actual change in lump-sum income less the additional income necessary to compensate the consumer, measured at the final utility level.

Jorgenson's Social Expenditure Function

Now return to Jorgenson's problem of constructing an appropriate income measure of the change in social welfare. Consider two social states, 0 and 1, defined by the vectors of prices, households' incomes, and households' characteristics in each situation. The vectors of prices, incomes, and characteristics in turn determine a vector of indirect utilities and a level of social welfare in each situation, given Jorgenson's translog estimates of the indirect utility function and his social welfare function:

$$\left(\overrightarrow{P}_0, \overrightarrow{M}^h_0, \overrightarrow{A}^h_0\right) \Rightarrow \ln \overrightarrow{V}^h_0 \Rightarrow W^0$$
$$\left(\overrightarrow{P}_1, \overrightarrow{M}^h_1, \overrightarrow{A}^h_1\right) \Rightarrow \ln \overrightarrow{V}^h_1 \Rightarrow W^1$$

The social expenditure function for each social state that corresponds to the individual consumer's expenditure function asks the question: What is the minimum aggregate level of (lump-sum) income required to achieve the actual level of social welfare in that social state? The problem is how to compute the minimum aggregate level of income. Consider social state 0, in which social welfare equals:

$$W^0 = \overline{\ln V_0} - g\left(\ln V_0 - \ln V^h_0; e\right)$$

Assume that there is some inequality ($V^h_0 \neq \overline{V}_0$, for some h) and some aversion to inequality ($e > 0$). Under these assumptions, the minimum aggregate level of income necessary to achieve W^0 must be less than the actual aggregate level of income in social state 0. The reason why is that social welfare would be maximized for the given sum of indirect utilities if everyone's utility were equal to the mean, $\overline{V}_0$. With no inequality, $g = 0$ and $W^0 = \ln \overline{V}_0$. The task, then, is to compute for each household the amount of income that would place the household at the mean level of utility, $W^0 = \overline{\ln V_0}$, given the actual prices and the household's characteristic(s) in social state 0: $\overrightarrow{P}_0$ and $\overrightarrow{A}^h_0$.

This can be done with the estimated translog utility function for each household, Eqn (4.28), reproduced here as Eqn (4.35):

$$\ln V^h = \ln p'\alpha_p + 1/2 \ln p'\beta_{pp} \ln P - D(P)\ln\left[M^h/m^0\left(P, A^h\right)\right] \tag{4.35}$$

Invert the utility function to represent $\ln M^h$ as a function of prices, the household's characteristic(s), and the household's utility. The inversion is possible because of the exact aggregation restrictions defined above, which make $\ln V^h$ linear in $\ln M^h$:

$$\ln M^h = 1/D(P)\left[\ln p'\alpha_p + 1/2 \ln p'\beta_{pp} \ln P - \ln V^h\right] + \ln m^0\left(P, A^h\right) \tag{4.36}$$

Finally, substitute $W^0 = \overline{\ln V_0}$ for $\ln V^h$ in Eqn (4.36), along with $\overrightarrow{P}_0$ and A^h_0, to find the level of income required to place the household at the mean level of utility. Call the required income M^h_*. M^h_* is less than M^h_{act} for those whose utilities are above the mean and greater than M^h_{act} for those whose utilities are less than the mean. Compute M^h_* for each household. The aggregate social expenditure (income) associated with W^0 is $\Sigma_h M^h_*$. Also, $\sum_h M^h_*$ must be less than $\Sigma_h M^h_{\text{act}}$ because the estimated indirect utility functions exhibit diminishing marginal utility of income under the exact aggregation assumption. With $\ln V^h$ linear in $\ln M^h$, V^h is concave in M^h. Therefore, when placing everyone at the mean level of utility the sum of the incomes taken away from those whose utilities are above the mean exceeds the sum of the incomes given to those below the mean.

Figure 4.7 illustrates the case of two households. V^h_H and V^h_L are equidistant from the mean, $\overline{V}$, but much more income must be taken from the high-utility person than must be given to the low-utility household to bring each to $\overline{V}$. Therefore, the aggregate income required to bring each to the mean is less than the actual aggregate income when the utilities are V^h_H and V^h_L.[9]

The minimum social expenditure $\sum_h M^h_*$ associated with social state 0 is a function of the prices, households' characteristics, and the level of social welfare in that state (the level of social welfare implicitly incorporates the

9. An alternative and instructive evaluation of the social expenditure function relies on the property that individual utilities are equalized if expenditures per household equivalent member, $M^h/m^0(P, A^h_0)$, are equalized. With expenditures per household equivalent member equalized,

$$\overline{\ln V} = \ln P'\alpha_p + 1/2B_{pp} \ln P - D(P)\ln\left[M/\sum m_0\left(P, A^h\right)\right] = W, \tag{4.36a}$$

where M is aggregate expenditures. Inverting the equation yields the log of the required minimum aggregate expenditures to achieve W:

$$\ln M = 1/D(P)\left[\ln p'\alpha_p + 1/2B_{pp} \ln P - W\right] + \ln\left[\sum m^0\left(P, A^h\right)\right] \tag{4.36b}$$

The aggregate social expenditure function depends on prices, the level of social welfare W, and the number of household equivalent members. Furthermore, the log of minimum aggregate social expenditures per capita is

$$\ln\left[M\Big/\sum m^0\left(P, A^h\right)\right] = 1/D(P)\left[\ln p'\alpha_p + 1/2B_{pp} \ln P - W\right] \tag{4.36c}$$

where the per capita measure is defined in terms of household equivalent members.

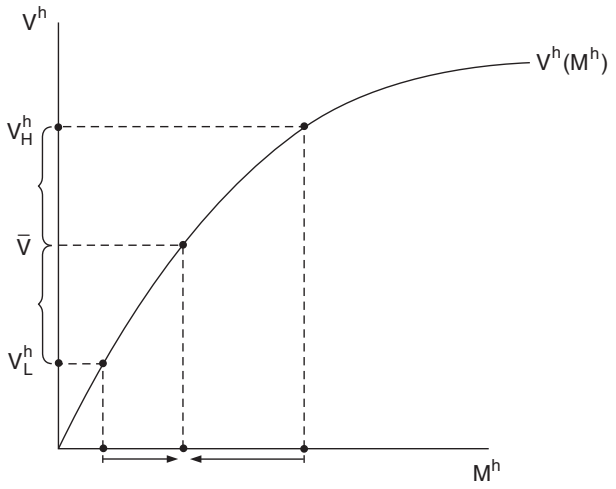

FIGURE 4.7

distribution of utilities and society's aversion to inequality). Therefore, write the social expenditure function for social state 0 as

$$M^0_{\text{soc}} = M^0\left(\overrightarrow{P}_0, \overrightarrow{A}^h_0, W^0\right) \tag{4.37}$$

The minimum aggregate income associated with social state 1 is derived in the same manner, using, $\overrightarrow{P}_1$, $\overrightarrow{A}^h_1$, and $W^1 = \overline{\ln V}^1$. Therefore, write the social expenditure function for social state 1 as

$$M^1_{\text{soc}} = M^1\left(\overrightarrow{P}_1, \overrightarrow{A}^h_1, W^1\right) \tag{4.38}$$

In general, a social expenditure function can be defined for any vector of prices, household characteristics, and a given level of social welfare, just as the consumer's expenditure function can be defined for any vector of prices and a given level of utility:

$$M_{\text{soc}} = M\left(\overrightarrow{P}, \overrightarrow{A}^h, W\right) \tag{4.39}$$

Social HCV and HEV

The social analogs to the individual HCV and HEV income measures of gains and losses follow naturally from the social expenditure function. Comparing a move from social state 0 to social state 1,

$$\text{HCV}_{\text{soc}} = M\left(\overrightarrow{P}_1, \overrightarrow{A}^h_1, W^0\right) - M\left(\overrightarrow{P}_1, \overrightarrow{A}^h_1, W^1\right) \tag{4.40}$$

The social HCV is the difference between the minimum aggregate income that would be required to achieve the original level of social welfare at the new prices and new household characteristics and the minimum aggregate income that would be required to achieve the new level of social welfare at the new prices and new household characteristics. If prices rose on average or household characteristics changed in such a way as to make households worse off, the HCV_{soc} would be positive. Society would have to receive a gift of income to maintain the level of social welfare:

$$\text{HEV}_{\text{soc}} = M\left(\overrightarrow{P}_0, \overrightarrow{A}^h_0, W^1\right) - M\left(\overrightarrow{P}_0, \overrightarrow{A}^h_0, W^0\right) \tag{4.41}$$

The social HEV is the difference between the minimum aggregate income that would be required to achieve the new level of social welfare at the original prices and original household characteristics and the minimum aggregate income that would be required to achieve the original level of social welfare at the original prices and original household characteristics. If prices rose on average or household characteristics changed in a way to make households worse off, the HEV_{soc} would be negative. Society would be willing to sacrifice some income to return to the original prices and household characteristics.

Two Applications for the US Economy

The US Standard of Living

Jorgenson estimated a system of five share equations for the United States from 1947 to 1985 using data from the National Income and Product Accounts, the Current Populations Survey, and the 1972–73 Survey of Consumer Expenditures. The five expenditure categories were energy, food, other nondurable goods, capital services from consumer durables and housing, and consumer services. He chose five household characteristics: family size, age of head of household, region of residence, race, and type of residence (urban, rural). Based on the estimates, he computed the HEV_{soc} for the United States from 1947 to 1985, expressed on a per capita basis using household equivalent members to represent the US standard of living. The per capita HEV_{soc} slightly more than tripled during that period in real terms, with an average annual rate of growth of 2.92. In contrast, the average annual rate of growth of real income per capita, the conventional measure of standard of living, was a much more modest 2.07% from 1947 to 1985.

Jorgenson attributes the overly pessimistic bias of the conventional income measure to three sources (with the percentage of the overall bias in parenthesis):

1. The use of the consumer price index (CPI) to deflate income rather than the price index implied by the Jorgenson approach to measuring social welfare,

which Jorgenson calls the social cost of living index.[10] The estimated social cost of living index grew more slowly than the CPI during this period (34.1% of the overall bias).

2. The use of a straight head count in arriving at a per capita measure rather than household equivalent members. The household equivalent member measure assumes that the household is the decision-making unit and takes account of changes in household characteristics over time. The number of household equivalent members grew more slowly than the overall population during this period (17.6% of the overall bias).
3. Ignoring equity entirely. The distribution of estimated utilities across household equivalent members became more equal during this period (48.2% of the overall bias).

Poverty in the United States

The Jorgenson consumption-based approach to measuring social well-being also gives a more optimistic picture of the extent of poverty in the United States relative to the official Department of Commerce poverty count, which is based on income. Daniel Slesnick estimated virtually the same expenditure system as Jorgenson using 13 years of data from the Consumer Expenditure Surveys (1961/62, 1972, 1973, and 1980 through 1989). He had the same five expenditure categories but added a sixth household characteristic, the sex of the head of household, to the five characteristics listed above. He then used his estimates to compute a consumption-based poverty head count for each of the 13 years.[11] The poverty computation involves three steps.

The first step is to define a consumption-based poverty line. Slesnick chose to define his poverty line similarly to the method that the Department of Commerce chose to compute the official poverty line in 1964. The Department of Commerce determined the minimum income a family required to purchase a nutritionally adequate diet and then multiplied the food budget by three to arrive at the "official" poverty line level of income. The poverty line varies by family size and composition (but no other characteristics) and is adjusted annually for changes in the CPI.

Slesnick chose a reference family and noted how much the Department of Commerce said it would have to spend on food in 1964 to purchase a nutritionally adequate diet. His reference family had the following characteristics: four people; headed by a white male 25−34 years old; living in a nonfarm area in the Northeast. Call the vector of reference characteristics A^R. Using the food equation from the estimated demand system, Slesnick determined the total expenditures, M^z, that would be consistent with purchasing the nutritionally adequate diet at 1964 prices for the reference family with characteristics A^R. Then, using Eqn (4.35), he determined the utility level, V_Z^R, achieved by the reference family at 1964 prices, expenditures M^z, and characteristics A^R. V_Z^R is the poverty line level of utility for a reference family, and M^z is the consumption-based poverty line level of expenditures. V_Z^R is assumed to remain constant over time.

Slesnick rescaled M^Z to 1973 prices since 1964 was not a year in his data set. Therefore, his poverty line M^Z was the individual expenditure function (4.36) evaluated at 1973 prices, characteristics A^R, and utility level V_Z^R:

$$M^Z = M\left(\overrightarrow{P}_{73}, A^R, V_Z^R\right) \tag{4.42}$$

The next step is to compute the utility level achieved in each year by each family in his data set. The utility level for family h at time t, V_t^h, is determined by Eqn (4.35) evaluated at current year prices $\overrightarrow{P}_t$, current family income M_t^h, and current family characteristics A_t^h.

The final step is to ask how much total expenditure each family in each year would have required to achieve utility level V_t^h if it could consume at 1973 prices and if it had the reference family characteristics A^R. This expenditure level is given by the individual expenditure function (4.36) evaluated at 1973 prices $\overrightarrow{P}_{73}$, characteristics A^R, and utility level V_t^h:

$$M_t^h = M\left(\overrightarrow{P}_{73}, A^R, V_t^h\right) \tag{4.43}$$

M_t^h evaluated in this way standardizes both for changes in prices since 1973 and for the needs of families with characteristics different from A^R. The number of poor equals the number of families for which $M_t^h < M^Z$. Alternatively, the number of poor equals the number of families for which $V_t^h < V_Z^R$, the poverty line level of utility, which is constant over time.

Slesnick found that his consumption-based poverty count was lower than the official poverty count in all but 4 years of the sample period and was substantially lower by the end of the period. From 1981 through 1989, Slesnick's estimated poverty rate was approximately four percentage points below the official poverty rate each year. For

10. The social cost of living index is based on the notion of the potential level of social welfare attainable in a given year, equal to the level of social welfare if the aggregate income were distributed to equalize utilities. The social cost of living is the ratio of the expenditures required to reach the potential social welfare at current prices to the expenditures required to reach the potential social welfare at base-year prices. The expenditures at base-year prices can be computed from Eqn (4.36).

11. Slesnick (1993). See also the articles by Jorgenson (1998), Triest (1998).

example, in 1989, Slesnick estimated that 8.4% of all families were poor, whereas the official poverty rate was 12.8%.

The consumption-based poverty rate is below the official income poverty rate primarily for three reasons. One is that 40% of the poor own their own homes, so that they consume a fairly large amount of capital services; capital service flows account for 10−13% of the total expenditures of the poor in Slesnick's sample. The second is that a large number of the poor dissave; their incomes are temporarily low and they dissave to maintain their standard of living. The third factor that drove the Slesnick poverty counts down sharply in the 1980s was a change in family characteristics that helped to move families out of poverty, both directly and indirectly by its effect on the composition of family expenditures.

Slesnick argues that his consumption-based poverty count is superior to the official count because it better reflects families' permanent economic situations. The poor by his measure are more likely to be permanent income poor than the "official" current income poor. He found that budgets of the consumption poor contain a lower percentage of capital services than the consumption nonpoor because they are less likely to own their own homes. He also found that the consumption poor devote a higher percentage of their budgets to purchases of food.

SOCIAL WELFARE AND SOCIAL MOBILITY

Bergson and Samuelson conceived of their individualistic social welfare function in terms of end-results equity, as a device for evaluating the ethical content of social outcomes. All our applications of the social welfare function so far have been in this vein. Despite its end-results orientation, economists have also used the social welfare function to measure the ethical implications of one common measure of process equity, the degree of social mobility in society.

Social mobility refers to the ability of individuals (families) to move throughout the distribution of income over time. It is closely related to the other widely held notion of process equity, equal opportunity. At one extreme is the caste system, a completely immobile society. People are assigned a position in the distribution at birth and can never move; there is no opportunity for change, much less equal opportunity. At the other extreme is complete mobility, in which people at any point on the distribution have an equal probability of staying there or moving to any other point on the distribution. A completely mobile society would almost certainly have full equality of opportunity along every relevant economic dimension.

The degree of social mobility is described by a transition probability matrix, defined as follows. Divide the income distribution into a number of categories (say, three for the purposes of illustration): low, $0 < Y_{\text{low}} \leq Y_1$; middle, $Y_1 < Y_{\text{middle}} \leq Y_2$; and high, $Y_{\text{high}} > Y_2$. Collect data on the position of the individuals (families) at time t and on the position of the same individuals (families) at time $t+1$, where $t+1$ may be 5−10 years beyond t. On the basis of these data, compute the 3×3 probability transition matrix:

$$P = [p_{ij}] \tag{4.44}$$

Each element, p_{ij}, is the probability that an individual (family) who was in income category i at time t is in income category j at time $t+1$.

Social Mobility and the Distribution of Income

The idea that movement through the distribution over time is governed by the transition probability matrix leads to a dramatic and well-known theorem. Assume that the matrix has the following three properties:

1. The p_{ij} are constant over time.[12]
2. $p_{ij} > 0$, for all i, j. There is always some probability that a person can move to any point on the distribution from any other point. Movement between two categories is never impossible, as it would be in a caste system.
3. The transition between income categories over time is a Markov process. The probability of a person being in income category j at time $t+1$ depends only on that person's position in time t. All history before time t is irrelevant to the distribution in time $t+1$.

These three assumptions are almost universally employed in the analysis of social mobility. They imply that the economy will eventually reach the same steady-state distribution of income regardless of the initial distribution of income.

The proof of this result is straightforward. Define the distribution vector $\pi'_t = (\pi^t_1, \pi^t_2, \pi^t_3)$, where π^t_i is the proportion (or number) of people in income category i at time t. Under the Markov assumption,[13]

$$\pi'_{t+1} = \pi'_t P \tag{4.45}$$

12. This is a truly heroic assumption given that the p_{ij} are influenced by so many factors, such as labor supply and saving behavior, trends in individual labor and capital markets, education decisions and markets, marriage patterns, social contacts, discrimination, and so forth.

13. For example, the first term in the multiplication on the RHS of Eqn (4.45) is $\pi^t_1 p_{11} + \pi^t_2 p_{21} + \pi^t_3 p_{31}$, equal to the sum of the proportion of people in the first category at time t times the probability that they stay in the first category, plus the proportion of people in the second category times the probability that they move to the first category, plus the proportion of the people in the third category times the probability that they move to the first category. The sum equals π^{t+1}_1.

Adding the other two assumptions, the steady-state distribution vector π is the solution to the system of equations:

$$\pi' = \pi' P \tag{4.46}$$

or

$$\pi'(I - P) = 0 \tag{4.47}$$

which has a unique solution for π' because $(I-P)$ is singular.

The intuition behind the result is that the spreading effect of P eventually dominates any initial distribution. Suppose the distribution is in the steady state at time $t-1$. Then, at time t, the government levels everyone to the mean with lump-sum taxes and transfers, in accordance with the first-best interpersonal equity conditions under the assumptions of equal marginal social welfare weights, identical tastes, and diminishing marginal utility of income. Everyone is now in middle income category 2 at t. By time $t+1$, however, some people will have moved to the other two income categories, the numbers determined by the probabilities p_{21} and p_{23}. In time $t+2$, the distribution will spread some more, as movement now occurs from all three income categories. The spreading continues until the original steady-state distribution of time $t-1$ is eventually reestablished.

The theorem points to a sharp tension between the process equity goal of social mobility and end-results equity goal of distributive justice. It implies that any redistribution of income undertaken in the name of end-results equity is ultimately futile. The underlying social mobility in the economy generated by P always returns the economy to the original steady-state distribution.

This tension is tempered by two considerations, however. One is that the government's redistribution policies will change the distribution until the economy returns to the steady state, and the new distributions during the transition periods may be social welfare increasing. The second is that any substantial redistribution of income will almost certainly change some of the elements of P. For instance, a complete leveling of the distribution would at the very least change people's labor supply and saving behavior. Whether the resulting changes in the p_{ij} are desirable, however, is another matter.

These considerations notwithstanding, the idea that the social mobility in the economy tends to undermine the government's redistribution policies strikes at one of the foundations of normative public sector theory, the first-best interpersonal equity conditions. The government may not be able to achieve the distribution implied by the interpersonal equity conditions of social welfare maximization as a steady-state distribution even in a first-best policy environment.

Structural Mobility, Circulation Mobility, and Social Welfare

The question remains whether the degree of social mobility itself has any direct bearing on social welfare as measured by the Bergson–Samuelson social welfare function. The answer is yes. Two features of the transition probability matrix P are related to social welfare. One is the steady-state distribution vector implied by P, which is commonly referred to as the structural mobility of the economy. Structural mobility is an element of end-results equity and, as such, has an obvious effect on social welfare. The other is the transition of the economy from any given distribution to its steady state, which is commonly referred to as the circulation mobility of the economy. The circulation mobility is the pure process equity component of P.

The limits of circulation mobility are given by the transition matrices:

1. $P = I$, the identity matrix
2. $P = [1/n]$, with $p_{ij} = 1/n$ for all i, j, and $n =$ the number of income categories

$P = I$ is the case of no circulation, the caste system. The distribution can never change because the given, initial distribution is the steady-state distribution. $P = [1/n]$ is the case of full circulation. From any initial distribution of income, the economy moves in one period to the steady-state distribution with an equal number of people in each income category.[14]

Valentino Dardanoni has provided an extensive analysis of the social welfare implications of circulation mobility Dardanoni (1993). We will highlight two of his main results, which relate to the question of whether circulation mobility has an independent effect on social welfare.

To focus on circulation mobility per se, Dardanoni begins by considering the set of transition probability matrices that have the same steady-state distribution. Two transition matrices P and Q have the same steady-state distribution if

$$\pi' = \pi' P = \pi' Q \tag{4.48}$$

This restriction is not very limiting because transition probability matrices that generate the same steady-state distribution can have very different transitional properties. For example, even the extreme transition matrices $P = I$ and $P = [1/n]$ (no circulation and full circulation) have the same steady state when the initial distribution is $\pi' = (1/n, \ldots, 1/n)$.

14. For example, with $n = 3$, $\pi_1^{t+1} = \pi_1^t(1/3) + \pi_2^t(1/3) + \pi_1^{3t}(1/3) = 1/3$ and likewise for π_2^{t+1} and π_3^{t+1}.

Dardanoni then argues that the appropriate arguments of a Bergson–Samuelson social welfare function are the expected discounted lifetime utilities of every individual. Define u_i as the utility received in any time period by people in income category i, the instantaneous utility. Assume that all people in income category i receive utility u_i, that u_i increases with income, and that, for simplicity, u_i remains constant over time. Identify people by the utility they receive in the initial distribution of income: A u_i person is someone in category i in the initial distribution. Define V_i as the expected discounted lifetime utility of a u_i person. Then, in matrix notation, the vector of expected discounted lifetime utilities V^P under the transition probability matrix P equals

$$V^p = u + \rho Pu + \rho^2 P^2 u + , \ldots , + \rho^n P^n u \tag{4.49}$$

where $\rho = 1/(1 + r_{soc}) =$ the social discount factor applied to future utilities with r_{soc} equal to the social marginal rate of substitution, and u is the vector of instantaneous utilities.[15] In the limit,

$$V^P = [I - \rho P]^{-1} u \tag{4.50}$$

Dardanoni normalizes the vector V^P by the discount factor, so that

$$V^P = (1 - \rho)[I - \rho P]^{-1} u \tag{4.51}$$

Define the matrix

$$P(\rho) = (1 - \rho)[I - \rho P]^{-1} \tag{4.52}$$

so that

$$V^P = P(\rho)u \tag{4.53}$$

$P(\rho)$ is a lifetime transition probability matrix. Its elements, $p(\rho)_{ij}$, can be interpreted as the discounted lifetime probability of moving from initial category i to final category j.

Utilitarian Social Welfare and Circulation Mobility

The arguments of the Bergson–Samuelson social welfare function are the elements of V^p, which are in turn a function of ρ, P, u, and n. Suppose that social welfare is utilitarian, as is commonly assumed. Then, given the steady-state distribution $\pi' = (\pi_i)'$ as

$$W^P = \sum_{i=1}^{N} \pi_1 V_i^P = \pi' V^P = \pi' P(\rho) u \tag{4.54}$$

Consider another transition probability matrix, Q, with the same steady-state distribution $\pi' = (\pi_i')$. Then,

$$W^Q = \sum_{i=1}^{N} \pi_i V_i^Q = \pi' V^Q = \pi' Q(\rho) u \tag{4.55}$$

The first result relating to social welfare is that $W^P = W^Q$. Circulation mobility has no effect on social welfare if the social welfare function is utilitarian. The proof follows immediately from the derivation of the steady-state distribution and the fact that P and Q generate the same steady state:

$$\pi' P(\rho) = \pi' P = \pi' = \pi' Q = \pi' Q(\rho) \tag{4.56}$$

Therefore,

$$W^p = \pi' P(\rho) u = \pi' Q(\rho) u = W^Q \tag{4.57}$$

Utilitarianism is indifferent to circulation mobility because it only cares about aggregate lifetime expected utility. It is completely indifferent to the composition of that aggregate, both in the steady state and as the economy evolves to the steady state over time.

Weighted Social Welfare and Circulation Mobility

The indifference of the utilitarian social welfare function to circulation mobility had been known for some time.[16] Dardanoni's contribution was to show that circulation mobility does have an independent effect on social welfare if social welfare is a weighted sum of the expected lifetime utilities. He also developed an empirical test for determining which of two transition matrices P and Q that generate the same steady state distribution yields the larger social welfare because of its superior transitional properties.

15. For example, the first term of Pu is $p_{11}u_1 + p_{12}u_2 + p_{13}u_3$, which is the expected utility in period 2 of a u_1 person, a person who is initially in category 1 at the bottom of the distribution. The entry in the first row, first column of P^2 is $p_{11}p_{11} + p_{12}p_{21} + p_{13}p_{31}$. It shows every path that a u_1 person can take and be in category 1 two periods from now: remain in category 1 in both periods, move to category 2 in period 1 and then back to category 1 in period 2, and move to category 3 in period 1 and then back to category 1 in period 2. Multiplying this sum by u_1 gives the expected utility of these paths for a u_1 person. The element in the first row, second column of P^2 shows every path that a u_1 person can take and be in the second category in period 2, and the element in the first row, third column of P^2 shows every path that a u_1 person can take and be in the third category in period 2. Multiplying these elements by u_2 and u_3, respectively, indicates the period 2 expected utility of a u_1 person who takes these paths. Therefore, the multiplication of the first row of P^2 and u gives the expected utility of a u_1 person in period 2. Similarly, the elements in the first row of P^n indicate every possible path that a u_1 person can take to be in categories 1, 2, and 3, respectively, in period n. V_1 is the discounted sum of these period-by-period expected utilities of a u_1 person, the lifetime expected utility.

16. The result was first demonstrated in 1986 by Kanbur and Stiglitz (Kanbur and Stiglitz, 1986).

Define a nonincreasing vector of social welfare weights, $\lambda' = (\lambda_i)$. Dardanoni argues that a nonincreasing weighting scheme is the natural distributional assumption in the context of social mobility if society cares about the poor and the transition probability matrix is monotonic, as is also commonly assumed. P is monotonic if it exhibits stochastic dominance, in the sense that it is always better in expected value terms to start in a higher income category. For example, P would be monotonic if

$$P_{11} > P_{21} > P_{31}; P_{11} + P_{12} > P_{21} + P_{22}; P_{21} + P_{22} > P_{31} + P_{32}; \text{and so forth}$$

(The proof of his weighted social welfare result does not require monotonicity; some other results in his paper do, however.)

As before, consider two transition probability matrices P and Q that generate the same steady-state distribution vector $\pi' = (\pi_i)'$. Define II as the diagonal matrix with the steady-state proportions in each income category on the diagonal. The weighted social welfare under each transition matrix is

$$W^P = \sum_{i=1}^{N} \lambda_i \pi_i V_i^P = \lambda' \Pi P(\rho) u \tag{4.58}$$

and

$$W^P = \sum_{i=1}^{N} \lambda_i \pi_i V_i^Q = \lambda' \Pi Q(\rho) u \tag{4.59}$$

Dardanoni asks: Under what set of conditions is $W^p > W^Q$?

The necessary and sufficient conditions make use of the summation matrix T, which has ones on and above the diagonal and zeros below the diagonal:

$$T = \begin{vmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{vmatrix} \quad T' = \begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{vmatrix}$$

PT generates the cumulative sums of each row in P, the cumulative density function for each income category. For example, the first row of PT is p_{11}, $P_{11} + P_{12}$, $P_{11} + P_{12} + P_{13}$. Similarly, $T'P$ generates the cumulative sums of each column in P. Also,

$$T^{-1} = \begin{vmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{vmatrix}$$

Premultiplying a vector by T^{-1} takes the differences of successive terms except the last term, which retains its value. For example,

$$T^{-1}u = \begin{vmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} u_1 \\ u_2 \\ u_3 \end{vmatrix} = (u_1 - u_2, u_2 - u_3, u_3)$$

Postmultiplying a vector by $(T^{-1})'[= (T')^{-1}]$ produces the same result. For example,

$$\lambda(T^{-1})' = (\lambda_1, \lambda_2, \lambda_3) \begin{vmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{vmatrix} = (\lambda_1 - \lambda_2, \lambda_2 - \lambda_3, \lambda_3)$$

Using the matrix T, Dardanoni's main theorem on weighted social welfare is that $W(V^P, \lambda) - W(V^Q, \lambda) \geq 0$ if and only if $T' \prod [P(\rho) - Q(\rho)] T \leq 0$, for λ nonincreasing and u nondecreasing.

To show that the second relationship implies the first, rewrite the first relationship as

$$\lambda' \Pi P(\rho) u - \lambda' \Pi Q(\rho) u \geq 0 \tag{4.60}$$

or

$$\lambda' \Pi [P(\rho) - Q(\rho)] u \geq 0 \tag{4.61}$$

Insert $I = (T)'^{-1} T' = T T^{-1}$ into the LHS to produce

$$\lambda' (T)'^{-1} T' \Pi \left[P(\rho) - Q(\rho)\right] T T^{-1} u \geq 0 \tag{4.62}$$

Consider the terms $T' \Pi [P(\rho) - Q(\rho)] T$ for the 3×3 case to illustrate the following properties:

1. The last row of the expression is zero. The last row of $T' \Pi$ is $\pi' = (\pi_1, \pi_2, \pi_3)$, the steady-state distribution vector. But $\pi' P(\rho) = \pi' = \pi' Q(\rho)$. Therefore, the last row of the expression is zero.
2. The last column of the expression is also zero. The last column of the matrix T sums the rows of $P(\rho)$ and $Q(\rho)$, both of which have to add to 1. Therefore, the last column of the expression is zero.

Next, consider the first two terms and last two terms of Eqn (4.62).

3. The first two elements of $\lambda' (T)'^{-1} = (\lambda_1 - \lambda_2, \lambda_2 - \lambda_3, \lambda_3)$ are ≥ 0.[17]
4. The first two elements of $T^{-1}u = (u_1 - u_2, u_2 - u_3, u_3)$ are ≤ 0.

Therefore, the entire expression is positive if $T' \Pi [P(\rho) - Q(\rho)] T \leq 0$, the sufficient condition for $W(V^P, \lambda) - W(V^Q, \lambda) \geq 0$.

17. Notice that the equal weights of the utilitarian social welfare function implies that the expression is zero, so that $W^P = W^Q$.

To show that $W(V^P, \lambda) - W(V^Q, \lambda) \leq 0$ implies $T'\Pi[P(\rho) - Q(\rho)]T \leq 0$, suppose to the contrary that the ij^{th} element of the second expression is positive. The difference in social welfare is given by Eqn (4.62). To establish a contradiction, select the vector λ such that it has ones for its first i elements and zeros thereafter, and select the vector u such that it has zeros for its first j elements and ones thereafter. Then, $\lambda'(T)'^{-1}$ has a 1 in the ith element and zeros everywhere else, and T^{-1} u has $a-1$ in its j^{th} element and last element and zeros everywhere else (the last element is unimportant). Having selected λ and u this way, the entire expression Eqn (4.62) is negative if the ij^{th} element of the second expression above is positive, a contradiction of $W(V^P, \lambda) - W(V^Q, \lambda) \geq 0$.

Finally, the expression $T'\Pi[P(\rho) - Q(\rho)]T \leq 0$ has a satisfying interpretation in terms of the social welfare implications of circulation mobility. Return to the 3×3 case for purposes of illustration and consider the first two rows and columns of $T'\Pi P(\rho)T$, which are the nonzero rows and columns in the entire expression. Postmultiplying $\Pi P(\rho)$ by T yields the cumulative sums of the rows, and then premultiplying by T' yields the cumulative sums of the columns of the cumulative row sums:

$$T'\Pi P\ (\rho)T = \begin{vmatrix} \lambda_1 p_{11} & (\lambda_1 p_{11} + \lambda_1 p_{12}) & \cdots \\ (\lambda_1 p_{11} + \lambda_2 p_{21}) & (\lambda_1 p_{11} + \lambda_1 p_{12} + \lambda_2 p_{21} + \lambda_2 p_{21}) & \cdots \\ \cdots & \cdots & \cdots \end{vmatrix}$$

For the entire expression to be negative, the corresponding elements in $T'\Pi Q(\rho)T$ must each be larger than the elements in $T'\Pi P(\rho)T$. Therefore, $Q(\rho)$ has less circulation mobility, and lower social welfare in the following sense: Individuals who start in category k or lower have a higher discounted probability of winding up, lifetime, in category j or lower for all k and j. In the expression $T'\Pi P(\rho)T$ (and $T'\Pi Q(\rho)T$), the row indicates the starting position and the column the ending lifetime position. Therefore, in the 3×3 case above, the first row, first column compares the probabilities of those who start and end in category 1. The first row, second column compares the probabilities of those who start in category 1 and end in either category 1 or 2. The second row, first column compares the probabilities of those who start in category 1 or 2 and end in category 1. And the second row, second column compares the probabilities of those who start in either category 1 or 2 and end in either category 1 or 2. These probabilities are all higher for $Q(\rho)$ if $Q(\rho)$ has lower social welfare than $P(\rho)$. Notice that the difference in welfare is entirely due to the difference in circulation mobility, in process equity, because $Q(\rho)$ and $P(\rho)$ both generate the same steady-state distribution. They have the same structural mobility, the same end-results equity.

In summary, Dardanoni has shown that an increase in upward mobility improves social welfare, but only if the social welfare function favors those with lower incomes.

Social Mobility in the United States

Thomas Hungerford measured the amount of social mobility in the United States by computing transition probability matrices over two 7-year time periods, 1969 to 1976 and 1979 to 1986. He divided the population into 10 income categories each time. His results were essentially the same in the two periods and tended to undercut the notion that the United States is the land of equal opportunity. The transition matrices were much closer to the no-circulation identity matrix than to the matrix of full circulation, despite the fairly fine gradation of the income categories. Most people stayed at or near to their original position in the distribution over a 7-year period. The p_{ij} declined sharply to very low levels at three or more deciles away from the initial position, at all points in the distribution. Hungerford concludes that there is not very much social mobility in the United States—bad news for process equity.

An offsetting piece of good news in Hungerford's data relates to end-results equity. Redistributional policies are not so quickly undermined by social mobility when the degree of social mobility is small. After all, if the transition probability matrix were the identity matrix, the redistribution would stick forever. Hungerford's data suggest that a social-welfare-improving redistributional policy may retain much of its impact for a very long time.[18]

REFERENCES

Amiel, Y., Creedy, J., Hurn, S., March 1999. Measuring attitudes towards inequality. Scandinavian Journal of Economics 101 (1), 83–96.

Atkinson, A., 1970. On the measurement of inequality. Journal of Economic Theory 2, 244–263.

18. See Hungerford (1993). See also Buchinsky and Hunt (1999). They found that young wage earners experienced lower social mobility and greater within-category inequality from 1979 to 1991.

Atkinson, A., 1983. The Economics of Inequality, second ed. Oxford University Press, New York. sect. 3.4, pp. 53–59.

Bishop, J., Formby, J., Smith, W., February 1991. Lorenz dominance and welfare: changes in the U.S. Distribution of income, 1967–86. Review of Economics and Statistics 73 (1), 134–139.

Blackorby, C., Donaldson, D., October 1988. Money metric utility: a harmless normalization? Journal of Economic Theory 46 (1), 120–129.

Buchinsky, M., Hunt, J., August 1999. Wage mobility in the United States. Review of Economics and Statistics 81 (3), 353–368.

Dardanoni, V., 1993. Measuring social mobility. Journal of Economic Theory 61, 372–394.

Fisher, F., 1987. Household equivalence scales and interpersonal comparisons. Review of Economics Studies LIV 54 (3), 519–524.

Harberger, A., 1983. Basic needs versus distributional weights in social cost-benefit analysis. In: Haveman, R., Margolis, J. (Eds.), Public Expenditure and Policy Analysis, third ed. Houghton Mifflin, Boston.

Hungerford, T., December 1993. U.S. Income mobility in the seventies and eighties. Review of Income and Wealth 39 (4), 403–417.

Jorgenson, D., September 1990. Aggregate consumer behavior and the measurement of social welfare. Econometrica 58 (5), 1007–1040.

Jorgenson, D., Spring 1985. Efficiency versus equity in economic policy analysis. American Economist 29 (1), 5–14.

Jorgenson, D., Winter 1998. Did we lose the war on poverty? Journal of Economic Perspectives 12 (1), 79–96.

Kanbur, S., Stiglitz, J., 1986. Intergenerational Mobility and Dynastic Inequality. Woodrow Wilson Discussion Paper No. 111, Princeton University.

Lambert, P., September 1993. Estimating impact effects of tax reforms. Journal of Economic Surveys 7 (3), 205–242.

Okun, A., 1975. Equality and Efficiency, the Big Tradeoff. The Brookings Institution, Washington, D.C.

Sen, A., 1982. Ethical measurement of inequality: some difficulties. In: Sen, A. (Ed.), Choice, Welfare, and Measurement. MIT Press, Cambridge, MA.

Shorrocks, A., 1983. Ranking income distributions. Economica 50 (197), 3–17.

Slesnick, D., February 1993. Gaining ground: poverty in the Postwar United States. Journal Political Economy 101 (1), 1–38.

Slesnick, D., December 1998. Empirical approaches to the measurement of welfare. Journal Economic Literature 36 (4), 2108–2165.

Triest, R., Winter 1998. Has poverty gotten worse? Journal of Economic Perspectives 12 (1), 97–114.

U.S. Census Bureau, March 2013. Current Population Survey. Historical Income Tables F-2 and H-2. http://www.census.gov/hhes/www/income/data/historical/index.html.

Chapter 5

The Problem of Externalities—An Overview

Chapter Outline

We begin our study of public expenditure theory with an analysis of externalities, which are a major source of inefficiency in any economy, market or otherwise. Externalities are often loosely defined as third-party effects, meaning that some activity by a set of economic agents affects other economic agents, "third parties," who are not directly engaged in the activity. This common definition is not precise enough for policy analysis, however. Because an economy is a highly interdependent system, almost any (important) economic activity generates repercussions—third-party effects—throughout the entire economy. Yet, not all economic activity requires public sector intervention.

POLICY-RELEVANT EXTERNALITIES

Consider the following two examples of externalities:

1. In the middle of the twentieth century, the demand for long-distance passenger travel shifted toward the airplane at the expense of the railroads.
2. A family living on the top of a hill builds a high fence around its property, which restricts the view previously enjoyed by many of its neighbors.

The first event triggered a huge number of third-party effects as the economy worked to accommodate the shift in demand. Generally speaking, resources specific to air travel gained, and those specific to rail travel lost, signaling a shift of resources away from the railroads and toward the airlines. Since people's tastes presumably differ, and different people received different incomes than before the shift to air travel, the whole pattern of demands for all goods and services tended to shift as well. These changes in demand occasioned still further changes in incomes and additional resource shifts to and from industries that may have been totally unrelated to air or rail travel, and so on, endlessly. Yet, the government did not necessarily have to intervene in this process. To the contrary, the very strength of the competitive market system is its ability to coordinate shifts in demands and resources, with changes in prices and profits acting as the signals that bring the economy to a new, efficient equilibrium.

In the second event, however, the third-party effects occur outside the normal market process. There is no natural market mechanism for recording the loss that each neighbor suffers from the fence. Any redress the neighbors might seek would presumably occur through the judicial process.

There is a second crucial difference in these two examples. In the first situation, the demand shifts in and of themselves have no effect on any of the fundamental *technical* relationships in the economy: the consumers' utility functions and the producers' production functions. All third-party gains and losses accrue through changes in prices, both goods and factor prices. Some consumers faced new budget constraints and some firms new profit functions, with corresponding gains or losses, all caused by the competitive process of supply and demand, which continuously changes consumer and producer prices while searching for a new equilibrium. In the second situation, in contrast, the neighbors lose because the properties of one of the variables in their utility functions, their land, have been altered and not because prices have changed. Each

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00005-4

neighbor's ability to enjoy his own property has diminished because of the fence, independently of any price changes generated by building the fence (of course, it is unlikely that prices would change in this case).

These two distinctions are the vital ones for public sector analysis. An externality, or third-party effect, may require government intervention to maintain efficiency if two conditions hold:

1. An activity by a set of economic agents enters ("alters") the utility functions of other consumers or the production functions of other producers not directly involved with the activity.
2. The gains and losses from these effects are not properly reflected in the competitive market system. This second condition is redundant in most cases, since externalities satisfying the first condition are almost never accounted for properly by the competitive market system.

Given the existence of externalities with these two properties, a perfect competitive market economy no longer generates a pareto-optimal allocation of resources. Government intervention may be required to keep society on its first-best utility−possibilities frontier. The only other possibility is private bargaining among the affected parties, which can be pareto optimal under certain conditions.

The Terminology of Externalities

Public sector economists struggled for years trying to pinpoint what kinds of third-party effects required government intervention. The puzzle was finally resolved in 1931 when Jacob Viner distinguished between pecuniary and technological externalities, terminology that remains in use today.[1] *Pecuniary externalities* refer to the market price effects illustrated by the first situation, those resulting directly from competitive market adjustments. They do not require public intervention to maintain pareto optimality. *Technological externalities* refer to third-party effects that satisfy the two conditions described above. These are the policy-relevant externalities.

The externality literature is filled with jargon to distinguish among the many different kinds of externalities. For instance, public sector economists distinguish between *external economies* and *diseconomies:* The former term refers to beneficial third-party effects, the latter to harmful third-party effects. Thus, one can speak of a "pecuniary external economy" or a "technological external diseconomy," and so forth. We will keep the distinction between economies and diseconomies in the text but drop the pecuniary/technological distinction. Because our only concern is for policy-relevant externalities, the term *externality* will always mean "technological externality" unless otherwise noted.

Another important distinction is among consumption, production, and consumption-production externalities:

Consumption Externality

Economic activity by some consumers enters (alters) the utility function of at least one other consumer but does not enter into (alter) any production relationships. The fence described above is an example of a consumer externality. Consumption of national defense is another more important example.

Production Externality

Economic activity by some firms enters (alters) the production function of at least one other firm but does not enter (alter) the utility function of any consumer. One firm removing oil from a common pool situated under land owned by more than one firm would be an example. The rate at which any one firm extracts the oil affects the total amount of oil that can be extracted from the pool by all the firms.

Consumption−Production Externality

Economic activity by some consumers enters (alters) the production function of at least one firm, or vice versa. Water pollution by a firm that affects both recreational and commercial fishing activities is an example of a consumption−production externality.

These distinctions are useful analytically because they generate different optimal policy rules. Chapters 6, 7, and 8 consider each of them in turn, beginning with consumption externalities in Chapter 6.

Still other terminological distinctions appear in the externality literature. We will develop them as needed within each chapter, whenever they are relevant for public policy.[2]

1. Viner (1952). The conceptual distinction was first noted by Allyn Young in 1913, but without Viner's terminology. Young (1913).

2. The treatment of externalities in Chapters 6−8 is comprehensive, with one notable exception. It does not consider an important type of externality called the *club good*, which was first analyzed by James Buchanan. A club good has the property that the extent of the externality can be controlled by the agents who generate the externality. For example, all members of a swim club have equal access to the club's swimming pool, but the club members control the total membership in the club. Buchanan's club good has appeared most prominently in the literature on fiscal federalism because a city or town can be viewed as a type of club. The standard economic model of a local jurisdiction assumes that only the citizens of a locality enjoy the public services offered by that locality, such as fire or police protection, and that the citizens determine the conditions of entry into the locality. We will hold off on presenting the club good until Part V on fiscal federalism. See Buchanan (1965).

THE ANALYSIS OF EXTERNALITIES: MODELING PRELIMINARIES

Chapter 3 described a useful property of first-best general equilibrium models that their first-order conditions dichotomize in two ways. One is that they generate distinct sets of interpersonal equity and pareto-optimal conditions. The former incorporate the social welfare function and describe how society can achieve end-results equity through lump-sum redistributions. The latter describe all the efficiency conditions necessary for society to achieve its utility–possibilities frontier. The pareto-optimal conditions do not contain any social welfare terms and can be achieved by competitive markets, absent any of the technical market failures such as externalities. The dichotomization of the interpersonal equity and pareto-optimal conditions was demonstrated in Chapter 2. The second dichotomy arises within the set of pareto-optimal conditions. Suppose that a technical market failure such as an externality exists in some markets. The externality changes the pareto-optimal conditions for that market, and government intervention may be required to achieve them. But the market failure has no effect on the form of the pareto-optimal conditions for all the other markets. Therefore, competitive markets can generate the pareto-optimal conditions in the unaffected markets; no government intervention is required in those markets. We will demonstrate the second dichotomy in Chapter 6.

These two dichotomies are useful because they permit formal analysis of policy problems with greatly condensed versions of the general equilibrium model presented in Chapter 2. For example, a consumer externality involves interrelationships among consumers only; producers are unaffected. Therefore, a first-best model analyzing a consumer externality can simply assume that production efficiency results from competitive markets, suppress the production side of the full model, and focus on the consumption externality among the consumers. Conversely, a production externality involves interrelationships among producers only. Therefore, a first-best analysis of a production externality can focus on the externality by positing a one-consumer equivalent economy, which assumes that competitive markets generate all the pareto-optimal conditions among consumers and that the government is optimally redistributing lump sum to satisfy the interpersonal equity conditions. These are legitimate assumptions in a first-best policy environment. Having analyzed the full model in Chapter 2, we know what the missing pareto-optimal and interpersonal equity conditions must be in the suppressed portions of the condensed models. Economists exploit these dichotomies all the time to analyze market failures with simple models. We will do the same throughout Part II.

Consider the following condensed version of the Chapter 2 model that is suitable for analyzing consumption externalities. The model deemphasizes production as much as possible while retaining all the essential consumption/utility elements from the full model. We will use it as our basic model in Chapter 6, adding only the particular external effects being analyzed. The condensed model is accomplished with the following modifications:

1. Define all goods and factors in terms of consumption by suppressing, notationally, the use of factors and the supply of goods by firms. Further, ignore the notational distinction between goods and factors, other than the convention that factors enter all utility and production relationships with a negative sign. Let

   X_{hi} = good i consumed by or factor i supplied by person h, $i = 1, \ldots, N$ and $h = 1, \ldots, H$.

   Notice that there are N total goods and factors in the economy (instead of the G goods and F factors in the model of Chapter 2).
2. Assume production is efficient and can be represented implicitly as a production–possibilities frontier in terms of the aggregate amount of consumer goods produced and factors supplied. Write

$$F(X_1, \ldots, X_i, \ldots, X_N) = 0 \tag{5.1}$$

   where X_i = the aggregate consumption (supply) of good (factor) i, and F( ) = an implicit function of all the relevant production relationships, corresponding to the production–possibilities frontier in two-good space.[3]
3. Finally, market clearance requires that

$$\sum_{h=1}^{H} X_{hi} = X_i \quad i = 1, \ldots, N$$

These constraints can be incorporated directly into the production–possibilities frontier, obtaining

$$F\left(\sum_{h=1}^{H} X_{h1}, \ldots, \sum_{h=1}^{H} X_{hi}, \ldots, \sum_{h=1}^{H} X_{hN}\right) = 0, \text{ or} \tag{5.2}$$

$$F\left(\sum_{h=1}^{H} X_{hi}\right) = 0 \tag{5.3}$$

3. Unless otherwise stated, we will always assume that F( ) describes a regular (convex outward) transformation surface for the economy. This in turn implies certain restrictions on the individual production functions. Kelvin Lancaster's *Mathematical Economics*, sections 8.4 through 8.7, contains an excellent analysis of the necessary and sufficient conditions on the individual production functions for a regular transformation surface. See Lancaster (1968).

with the understanding that producers do not care who receives (supplies) an additional unit of a good (factor). That is,

$$\frac{\partial F}{\partial X_{hi}} = \frac{\partial F}{\partial X_i} = F_i \quad \text{all } h = 1, \ldots, H$$

With these three condensations, the social welfare maximization problem becomes extremely simple to represent formally:

$$\max_{(X_{hi})} \; W[U^h(X_{hi})]$$
$$\text{s.t. } F\left(\sum_{h=1}^{H} X_{hi}\right) = 0$$

where W is the Bergson–Samuelson individualistic social welfare function.

Although this is a drastically condensed version of the original model, it is still perfectly valid as a general equilibrium model in a first-best environment. Furthermore, it is sufficiently general to generate all relevant pareto-optimal conditions involving consumption, as well as the standard interpersonal equity conditions. As such, it is ideal for analyzing consumer externalities, which essentially involves specifying which goods and factors enter whose utility functions.

The Interpersonal Equity Conditions

Consider first the interpersonal equity conditions. They are obtained from the first-order conditions with respect to any single good (or factor) consumed (supplied) by any two people, say X_{h1} and X_{j1}. Setting up the Lagrangian equation,

$$\max_{(X_{hi})} \; L = W\left[U^h(X_{hi})\right] + \lambda F\left(\sum_{h=1}^{H} X_{hi}\right)$$

and differentiating yields

$$X_{h1}: \frac{\partial L}{\partial X_{h1}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{h1}} + \lambda F_1 = 0 \tag{5.4}$$

$$X_{j1}: \frac{\partial L}{\partial X_{j1}} = \frac{\partial W}{\partial U^j}\frac{\partial U^j}{\partial X_{j1}} + \lambda F_1 = 0 \tag{5.5}$$

Therefore,

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{h1}} = -\lambda F_1 \quad \text{all } h = 1, \ldots, H \tag{5.6}$$

The social marginal utility of consumption of good 1 should be equalized across all people. This is the same rule obtained in the more detailed model of Chapter 2.

The Pareto-Optimal Conditions

To derive the pareto-optimal conditions for consumption, consider the first-order conditions with respect to two goods consumed by any one person, say, X_{hi} and X_{hk} (X_{hi} and X_{hk} could also be any two goods, any two factors, or any one good and any one factor).

$$X_{hi}: \frac{\partial L}{\partial X_{hi}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hi}} + \lambda F_i = 0 \tag{5.7}$$

$$X_{hk}: \frac{\partial L}{\partial X_{hk}} = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{hk}} + \lambda F_k = 0 \tag{5.8}$$

Rearranging, dividing, and simplifying,

$$\frac{\frac{\partial U^h}{\partial X_{hi}}}{\frac{\partial U^h}{\partial X_{hk}}} = \frac{F_i}{F_k} \quad \begin{matrix}\text{all } h = 1, \ldots, H \\ \text{any } i, k = 1, \ldots, N\end{matrix} \tag{5.9}$$

The ratio F_i/F_k gives the marginal rate of transformation (substitution) in production between goods (factors) i and k, and the left-hand side is their marginal rate of substitution in consumption. Hence, the single set of relationships, Eqn (5.9), reproduces pareto-optimal conditions P1, P2, P3, P6, P7, and P8 from the full model of Chapter 2 (recalling that i and k can be any two goods, any two factors, or any one good and any one factor). Only the production efficiency conditions P4 and P5 cannot be reproduced with this model, but they are assumed to hold whenever production is represented as an implicit production–possibilities frontier. Production must occur on the contract locus in factor space for the economy to be on its production–possibilities frontier.

Thus, condensed versions of the standard model such as this one retain a substantial amount of analytical flexibility despite their simplicity. This is why they are so useful for analyzing public sector problems in a first-best framework.

Chapter 6 turns to the analysis of consumption externalities using the condensed model. As we shall see, the analysis of any consumer externality requires only a simple modification of the condensed model. All one need to specify is which variables appear in each person's utility function.

REFERENCES

Buchanan, J., February 1965. An economic theory of clubs. Economica 32 (125), 1–14.

Lancaster, K., 1968. Mathematical Economics. Macmillan, New York.

Viner, J., 1952. Cost Curves and Supply Curves," Zeitschrift fur Nationalokonomie, III, 1932 reprinted in American Economic Association Readings in Price Theory. Richard D. Irwin, Chicago.

Young, A.A., August 1913. Pigou's Wealth and welfare. Quarterly Journal of Economics 24 (4), 672–686.

Chapter 6

Consumption Externalities

Chapter Outline

A policy-relevant consumption externality occurs whenever economic activity by some consumer enters (alters) the utility function of at least one other consumer and is not accounted for by the market system. The very definition itself suggests that the fundamental problem in analyzing consumption externalities is deciding exactly what activities enter whose utility functions and in precisely what form. Once the arguments of each consumer's utility function are specified, they determine every relevant feature of the consumer externality, including the government policy required to achieve pareto optimality.

We will make use of variations of the condensed general equilibrium model described in Chapter 5. Let X_{ik} represent the consumption (supply) of good (factor) k by person i, where

$$k = 1, \ldots, N \ (N \text{ total goods and factors})$$

$$i = 1, \ldots, H \ (H \text{ people})$$

Then the basic model for analyzing all consumption externality problems in a first-best policy environment is

$$\max_{(X_{ik})} W\left[U^h(\,)\right]$$

$$\text{s.t. } F\left(\sum_{i=1}^{H} X_{ik}\right) = 0$$

W is the individualistic social welfare function to be maximized and F is the implicit aggregate production possibility frontier. It assumes production efficiency and incorporates the market clearing equations for the N goods and factors. Also,

$$\frac{\partial F}{\partial X_{ik}} = \frac{\partial F}{\partial X_k} = F_k \quad \text{all } i = 1, \ldots, H_i; \quad \text{any } k = 1, \ldots, N$$

Producers do not care who consumes each good (supplies each factor). The nature of the consumer eternality

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00006-6

depends entirely on how the X_{ik} enter each person's utility function, the U^h.

HOW BAD CAN EXTERNALITIES BE?

Let us begin by considering the most intractable externality case and ask: How bad can consumption externalities be? The worst possible situation imaginable would require a triple indexing of X, X_{ik}^j, with X_{ik}^j entering the utility function of each person h:

$$U^h = U^h\left(X_{ik}^j\right)$$

X_{ik}^j refers to the consumption (supply) of good (factor) k by person i, affecting person j, $i = 1, \ldots, H$; $j = 1, \ldots, H$; and $k = 1, \ldots, N$. That is, each person h worries about who consumes (supplies) what good (factor) and how it affects each person.

Return to the example of the fence in Chapter 5, in which each person in a given neighborhood is affected whenever anyone builds a fence. Suppose there are H people in the neighborhood and person i builds a fence, good k. Each person h in the neighborhood notes that person i built the fence (good k) and that the fence affects everyone in the neighborhood differently. Thus, from the point of view of person h, X_{ik}^j is different from X_{ik}^ℓ, for $\ell \neq j$ and $\ell, j = 1, \ldots, H$. Each variable refers to person i's fence, but persons j and ℓ react differently to the fence and each person h in the neighborhood takes note of this difference. Had someone else built a fence (say, person m), then each person's utility function would contain another H argument, X_{mk}^j, $j = 1, \ldots, H$, and so forth. In the worst of all worlds, anything anyone did would affect everyone, and each person would take note of how everyone was affected by any one person's consumption of any good. Hence, each utility function would contain all H^2N elements, X_{ik}^j, as arguments. This would surely be the worst possible consumption externality situation imaginable.

Fortunately, we can at least dispense with the superscript j without disservice to any realistic situation. Continuing with the fence example, when any one person h considers the effects of the fence on himself and his $(H - 1)$ neighbors, we can assume that the H separate effects combine to generate a single overall effect on person h's utility. Thus, person h's utility function need only records that person i built a fence (good k), as opposed to someone else building a fence. At most, then, we need to place HN arguments, X_{ik}, in each person's utility function. Write

$$U^h = U^h\left(X_{ik}\right) \quad \text{any } h = 1, \ldots, H; \text{ all } i = 1, \ldots, H; \text{ and } k = 1, \ldots, N \tag{6.1}$$

to indicate that, in the worst of all worlds, each person h is affected by anyone's (i) consumption (supply) of any good (factor) k. The fact that person h considers the effects of some X_{ik} on all people is simply summarized as one effect on his utility, $U^h(X_{ik})$.

The general equilibrium social planner's model in this worst of all worlds becomes

$$\max_{(X_{ik})} W\left[U^h\left(X_{ik}\right)\right]$$
$$\text{s.t. } F\left(\sum_{i=1}^{H} X_{ik}\right) = 0$$

Notice that the goods (and factors) in this model are exclusive goods. X_{ik} means that person i physically consumes (supplies) good (factor) k, as indicated by the market clearance relationship $\sum_{i=1}^{H} X_{ik} = X_k$. X_{ik} enters into the utility function of all $(H - 1)$ other persons, but they are merely affected by X_{ik}; they do not physically consume it. Thus, there are $H \cdot (H - 1)$ *external* effects associated with the consumption (supply) of good (factor) k, and $H \cdot (H - 1) \cdot N$ total external effects, counting all N goods and factors in the worst of all possible worlds.

Externalities of this type are referred to as *individualized externalities* because the external effects depend on who is engaged in the exclusive activity that generates the externalities. It matters who builds the fence.

In the context of this model, a natural definition of a *pure public good (factor)* is

$$\frac{\partial U^h}{\partial X_{ik}} \neq 0 \quad \text{all } i, h = 1, \ldots, H \tag{6.2}$$

If everyone is affected *on the margin* by anyone's consumption (supply) of good (factor) k, then k is a pure public good. The choice of marginal rather than total utility in the definition makes sense because, as we shall see, it is marginal utilities (more precisely, marginal rates of substitution) that enter into the pareto-optimal decision rules. Person h could be significantly affected by person i's consumption of good k in a total sense, but if the marginal effect is zero, then it turns out that person h's feelings do not matter for purposes of allocational efficiency at the optimum.[1]

Note that our definition of publicness says nothing about the signs of $\partial U^h/\partial X_{ik}$. For some h, the derivative could be positive, for others negative, so long as $\partial U^h/\partial X_{ik}$ is never zero. The smoking of marijuana comes to mind as an example. Some people enjoy the fact that others indulge;

1. For the benefit of those somewhat familiar with the externality literature, we should also note that this definition differs from Samuelson's early definition of a pure public good, which has gained fairly wide acceptance. Samuelson equated publicness to nonexclusiveness or jointness in consumption, meaning that if any one person consumes the services of a good, then everyone automatically consumes their services. In our model, in contrast, only person i consumes X_{ik}, only person j consumes X_{jk}, and so forth, for any $i, j = 1, \ldots, H$, and X_{ik} does not necessarily equal X_{jk}. (For more on nonexclusive goods, refer to the next section of this chapter.)

other people clearly dislike it. In the terminology of externalities, marijuana generates both external economies and diseconomies.

Correspondingly, a *pure private good (factor)* is one for which

$$\frac{\partial U^h}{\partial X_{ik}} = 0 \quad i \neq h \tag{6.3}$$

Only person i is affected on the margin by his or her consumption (supply) of good (factor) k. We will write $U^h(X_{hk})$ to indicate that good (factor) k is a pure private good (factor), and $U^h(X_{ik})$ to indicate that a consumer externality exists that is *potentially* a pure public good. We say potentially because all the notations imply that some person h is affected by at least one other person's consumption (supply) of good (factor) k as well as his or her own consumption of good k. It is not meant to imply that everyone is necessarily affected by each person's consumption of good (factor) k. $\partial U^h/\partial X_{ik}$ could equal zero for some i or even most i. All that is required for the existence of a consumption externality is that one person's utility be a function of one other person's consumption (supply) of something.

THE WORST OF ALL WORLDS—ALL GOODS (FACTORS) ARE PURE PUBLIC GOODS (FACTORS)

In the worst of all worlds, all goods (factors) are pure public goods (factors).[2] For policy purposes, this is really a horrendous situation as the government can hardly interpret what the proper decision rules mean let alone have any hope of implementing them. The government's problem is

$$\max_{(X_{ik})} W\left[U^h\left(X_{ik}\right)\right]$$
$$\text{s.t. } F\left(\sum_{i=1}^{H} X_{ik}\right) = 0$$

with the understanding that each utility function $U^h(\,)$ contains all NH elements, X_{ik}, $i = 1, \ldots, H$; $k = 1, \ldots, N$.

The corresponding Lagrangian equation is

$$\max_{(X_{ik})} L = W\left[U^h\left(X_{ik}\right)\right] + \lambda F\left(\sum_{i=1}^{H} X_{ik}\right)$$

Before proceeding, notice how deceptively similar this problem is to the problem of social welfare maximization when there are only pure private goods. In our notation, the pure private goods case is represented as

$$\max_{(X_{hk})} W\left[U^h\left(X_{hk}\right)\right]$$
$$\text{s.t. } F\left(\sum_{h=1}^{H} X_{hk}\right) = 0$$

In each case, maximization occurs with respect to HN goods and factors, the N goods and factors consumed and supplied by each of H people. The difference is that in the worst of all worlds, all HN variables appear in each utility function, whereas in the pure private goods (factors) world, only N variables appear in each utility function. This difference matters, because the policy implications are enormously different. In the latter case, the competitive market can achieve full pareto optimality. In the former case, the market cannot be expected to achieve high efficiency, and the government is virtually powerless to act in an optimal manner. The problems for the government in the pure public goods case are self-evident upon examination of the first-order conditions for social welfare maximization, both the interpersonal equity conditions and the pareto-optimal conditions.

Interpersonal Equity Conditions

Recall that the interpersonal equity conditions are obtained by comparing the first-order conditions for any one good (factor) consumed (supplied) by any two persons, say X_{j1} and X_{i1}. The first-order conditions are

$$X_{j1}: \sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{j1}} = -\lambda F_1 \tag{6.4}$$

$$X_{i1}: \sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{i1}} = -\lambda F_1 \tag{6.5}$$

From conditions (6.4) and (6.5),

$$\sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{i1}} = -\lambda F_1 \quad \text{all } i = 1, \ldots, H \tag{6.6}$$

The interpretation of the interpersonal equity conditions is identical to that of the standard model in Chapter 2, which contained only pure private goods: The government should redistribute good 1, lump sum, until social welfare is equalized on the margin across all individuals. This task, difficult enough with pure private goods, is now hopelessly complex, however. When the government gives (takes) an extra unit of good (factor) 1 to (from) person i, it must know how all people react to that transfer (tax) on the margin, not just how person i's utility is affected, and similarly for units transferred to or from any other person. This is clearly an impossible task, one the government could not even hope to approximate.

2. Ng provides an alternative model of this worst of all worlds in Ng (1975).

Pareto-Optimal Conditions

The pareto-optimal conditions also differ considerably from their counterpart in a world of pure private goods, both in form and interpretation. Recall that the pareto-optimal conditions are obtained from the first-order conditions of any two goods consumed (factors supplied) by any one person, say X_{ik} and X_{il}. The first-order conditions are

$$X_{ik}: \sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{X_{ik}} = -\lambda F_k \tag{6.7}$$

$$X_{i1}: \sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{X_{i1}} = -\lambda F_1 \tag{6.8}$$

Dividing Eqn (6.7) by Eqn (6.8) yields

$$\frac{\sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{ik}}}{\sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{i1}}} = \frac{F_k}{F_1} \quad \begin{matrix} \text{all } i = 1, \ldots, H \\ \text{any } k = 2, \ldots, N \end{matrix} \tag{6.9}$$

The right-hand side (RHS) of Eqn (6.9) has a standard interpretation, the marginal rate of transformation (MRT) in production between goods (factors) k and 1. The left-hand side (LHS) has no standard interpretation, however. As written, it is a ratio of marginal impact on social welfare from consuming (supplying) the two goods (factors), and there is no way to simplify the expression. In particular, the social welfare terms, $\partial W/\partial U^h$, do not cancel, so that the rule is not really a pareto-optimal or efficiency condition at all. Recall that pareto-optimal conditions do not contain social welfare terms. In this worst of all worlds, then, the model does not dichotomize into interpersonal equity and pareto-optimal conditions, the only exception we will encounter in all of Part II. All the decision rules are of the interpersonal equity type and can be achieved only by lump-sum redistributions of all goods and factors, a truly hopeless situation. Moreover, the competitive market system, which equates marginal rates of substitution in consumption to marginal rates of transformation, would be absolutely useless. Nothing short of a complete government takeover of the economy would be capable of satisfying the first-order conditions for social welfare maximization, even in principle.

THE EXISTENCE OF AT LEAST ONE PURE PRIVATE GOOD

Fortunately, the real world is not so riddled with consumption externalities. A large number of goods are pure private goods, or close enough to pure private goods that a government would not consider intervening in their markets. To keep the discussion as general as possible, however, let us assume that there is only one pure private good in the economy, the first. Formally, $\partial U^h/\partial X_{i1} = 0$, $i \neq h$. The other $(N - 1)$ goods and factors remain pure public goods. As it turns out, only one private good is needed to resurrect the dichotomy between the pareto-optimal and interpersonal equity conditions, which normally exists in first-best analysis and to retain a role for the competitive market system in allocating all the goods and factors.

With a single private good, the social welfare maximization problem becomes

$$\max_{(X_{ik};\, X_{h1})} W\left[U^h\left(X_{ik}; X_{h1}\right)\right]$$
$$\text{s.t. } F\left(\sum_{i=1}^{H} X_{ik}; \sum_{h=1}^{H} X_{h1}\right) = 0$$

where $k = 2, \ldots, N$. Good 1 has been written separately to indicate specifically that it is a pure private good.

Interpersonal Equity Conditions

Consider the interpersonal equity conditions with respect to good 1, the pure private good. The first-order conditions are[3]

$$X_{h1}: \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{h1}} = -\lambda F_1 \tag{6.10}$$

$$X_{i1}: \frac{\partial W}{\partial U^i} \frac{\partial U^i}{\partial X_{i1}} = -\lambda F_1 \tag{6.11}$$

or

$$\frac{\partial W}{\partial U^h} \frac{\partial U^h}{X_{h1}} = -\lambda F_1 \quad \text{all } h = 1, \ldots, H \tag{6.12}$$

Equation (6.12) is identical to the interpersonal equity conditions in the standard model of Chapter 2. Assume that the government can redistribute X_1 lump sum to achieve this condition as part of its first-best policy strategy.

Pareto-Optimal Conditions

As above, consider the first-order conditions with respect to two goods (factors) consumed (supplied) by any one person i, say X_{ik} and X_{il}. The choice of k is arbitrary, but good 1, the private good, must be one of the two goods chosen. The first-order conditions are

$$X_{ik}: \sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{X_{ik}} = -\lambda F_k \tag{6.13}$$

$$X_{i1}: \frac{\partial W}{\partial U^i} \frac{\partial U^i}{\partial X_{i1}} = -\lambda F_1 \tag{6.14}$$

3. λ is the Lagrangian multiplier associated with $F(\,)$.

Dividing Eqn (6.13) by Eqn (6.14) yields

$$\frac{\sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{ik}}}{\frac{\partial W}{\partial U^i} \frac{\partial U^i}{\partial X_{i1}}} = \frac{F_k}{F_1}, \quad \text{for } k = 2, \ldots, N \tag{6.15}$$

Condition (6.15) can be simplified if the government has satisfied the interpersonal equity conditions for good 1. The LHS is a summation of social welfare terms over a common denominator, $\frac{\partial W}{\partial U^i} \frac{\partial U^i}{\partial X_{i1}}$. But, if interpersonal equity holds,

$$\frac{\partial W}{\partial U^i} \frac{\partial U^i}{\partial X_{i1}} = -\lambda F_1 \quad \text{all } i = 1, \ldots, H \tag{6.16}$$

Selectively substitute for the denominator term by term, matching up the social welfare terms, and write

$$\frac{\frac{\partial W}{\partial U^1} \frac{\partial U^1}{\partial X_{ik}}}{\frac{\partial W}{\partial U^1} \frac{\partial U^1}{\partial X_{11}}} + \ldots + \frac{\frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{ik}}}{\frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{h1}}} + \ldots + \frac{\frac{\partial W}{\partial U^H} \frac{\partial U^H}{\partial X_{ik}}}{\frac{\partial W}{\partial U^H} \frac{\partial U^H}{\partial X_{H1}}} = \frac{F_k}{F_1}, \quad \text{any } k = 2, \ldots, N \tag{6.17}$$

$$\sum_{h=1}^{H} \left[\frac{\frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{ik}}}{\frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{h1}}} \right] = \frac{F_k}{F_1} \quad \text{all } i = 1, \ldots, N \text{ any } k = 2, \ldots, N \tag{6.18}$$

The social welfare indexes, $\partial W/\partial U^h$, cancel term by term, yielding

$$\sum_{h=1}^{H} \left[\frac{\frac{\partial U^h}{\partial X_{ik}}}{\frac{\partial U^h}{\partial X_{h1}}} \right] = \frac{F_k}{F_1} \tag{6.19}$$

The LHS of Eqn (6.19) has a standard pareto-optimal interpretation, devoid of social welfare terms. It is a sum of marginal rates of substitution, each person's marginal rate of substitution (MRS) between person i's consumption of good k and his or her own consumption of the pure private good. Thus, the rule can be written as

$$\sum_{h=1}^{H} \text{MRS}^h_{X_{ik}, X_{h1}} = \text{MRT}_{k,1} \quad \text{for all } i = 1, \ldots, H \text{ any } k = 2, \ldots, N \tag{6.20}$$

Note carefully that the ability to cancel the social welfare terms is not just a formal "trick." It implies an optimal first-best policy action, a lump-sum redistribution that satisfies the interpersonal equity conditions for good (factor) 1. Without the optimal redistribution, the terms would not cancel and all the policy implications of the pareto-optimal conditions that we are about to discuss become irrelevant. Condition (6.19) would not be the necessary condition for a social welfare maximum. We will employ this cancellation technique repeatedly throughout this chapter, with the same policy implications understood each time. Without the ability to achieve correct lump-sum redistributions, none of the standard first-best policy prescriptions apply, even those ostensibly related only to allocational issues.[4]

Note, finally, that only good 1 need be redistributed lump sum, exactly as in the baseline private goods model of Chapter 2. If the government correctly redistributes good 1 and designs policies to achieve all the pareto-optimal conditions, then the interpersonal equity conditions automatically hold for goods (and factors) $k = 2, \ldots, N$ as well. To see this, plug the social welfare terms back into the LHS of Eqn (6.19), obtaining Eqn (6.18). If Eqn (6.18) holds and the denominators are also equal from interpersonal equity, then the numerators are also equal:

$$\sum_{h=1}^{H} \frac{\partial W}{\partial U^h} \frac{\partial U^h}{\partial X_{ik}} = -\lambda F_k \quad \text{all } i = 1, \ldots, H \text{ any } k = 2, \ldots, N \tag{6.21}$$

as required by the interpersonal equity conditions for goods $k = 2, \ldots, N$.

Because the pareto-optimal rules for externality-generating exclusive goods are combinations of marginal rates of substitution and marginal rates of transformation, they have the following properties:

1. One can always describe a market structure with competitive prices that will achieve the correct pareto-optimal conditions without government intervention. This is so because producers and consumers equate market prices to marginal rates of transformation and substitution under perfect competition. The necessary market structure is far more complex than the normal competitive market structure, however. It requires an entire new set of competitive market transactions among consumers that correctly accounts for all the external effects. In other words, the market failure associated with externalities can be thought of as a problem of nonexistent markets, namely the required competitive side markets among consumers.[5]
2. The government can achieve the pareto-optimal conditions within the standard decentralized competitive markets for each of the goods (factors) by levying a set of taxes or subsidies that directs competitive behavior to the correct pareto-optimal conditions.

These two properties are worth extended discussions.

4. Notice that condition (6.19) can be derived without reference to the social welfare function by solving the following problem: Maximize the utility of any one person, subject to holding the utilities of the remaining $(H - 1)$ people constant, and the production frontier and market clearance. But, the first-order conditions for this problem are not the necessary conditions for a social welfare maximum if the distribution is not optimal, in general. The first-order conditions for externalities with a nonoptimal distribution are derived in Chapter 22.

5. Kenneth Arrow argues for this view of the externality problem in Arrow (1977).

Externalities as Market Failure: The Missing Side Markets

The property that pareto-optimal decision rules for exclusive activities that generate externalities can always be achieved by an appropriate set of competitive markets follows directly from the assumptions of profit and utility maximization. Suppose, as above, that goods (factors) $k = 2, \ldots, N$ are pure public goods and that good 1 is a private good. Suppose, also, that the markets for all the goods (factors) are competitive, and that $P_1 = 1$ (good 1 is the numeraire). The standard competitive markets generate the conditions:

$$\text{MRS}^h_{X_{hk},X_{h1}} = \text{MRT}_{k,1} \quad \begin{matrix}\text{all } h = 1, \ldots, H \\ \text{any } k = 2, \ldots, N\end{matrix} \tag{6.22}$$

because both consumers and producers face the identical prices P_k and P_1 for goods (factors) k and 1, respectively. $\text{MRS}^h_{X_{hk},X_{h1}}$ refers to person h's MRS between his or her own consumption of goods k and 1 (supply of factors k and 1). These are not the pareto-optimal conditions given by Eqn (6.20). Additional competitive side markets are needed to achieve condition (6.20).

To understand the nature of these side markets, consider again person i's consumption of public good k, X_{ik} (assume both X_k and X_1 are goods). The competitive market structure that would generate the pareto-optimal condition,

$$\sum_{h=1}^{H} \text{MRS}^h_{Xik,X_{h1}} = \text{MRT}_{k,1}$$

is as follows. Producers insist on a price P_k, equal to $\text{MRT}_{k,1}$ (=MC_k with $P_1 = \text{MC}_1 = 1$), to supply good k. If consumer i wants to buy X_k, he or she has to pay the producer this price. Suppose X_{ik} generates an external economy (a "good") for all other consumers. In this case, person i and all the others have a mutual interest in developing side markets to influence the final value of X_{ik}, the others because they would be willing to pay something to have person i increase his or her consumption of good k, and person i because he or she can extract side payments that effectively lower the price to his or her below the producer price, P_k.

Consider next Fig. 6.1, which shows the set of indifference curves for some person $h \neq i$, between X_{ik}, person i's consumption of good k, and X_{h1}, person h's own consumption of good 1. X_{ik} is a parameter for person h, but he or she determines his or her own consumption of good 1. Suppose their independent decisions place consumer h at point B on indifference curve I_1. The slope of I_1 at B is $\text{MRS}^h_{X_{ik},X_{h1}}$. If these were two purely private goods both under the control of person h, then he or she would pay a competitive price for X_{ik} equal to $\text{MRS}^h_{X_{ik},X_{h1}}$. Call this price P^h_{ik} (with $P_1 \equiv 1$). Suppose person h actually paid person i

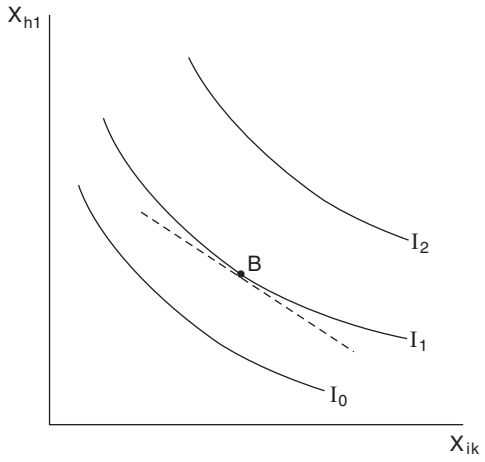

FIGURE 6.1

the competitive price $P^h_{ik} = \text{MRS}^h_{X_{ik},X_{h1}}$, and all other consumers did likewise, having formed identical "competitive" side market relationships with person i. This set of competitive side markets could achieve the desired pareto-optimal condition.

The effective price of good k to person i is

$$P^i_k = P_k - \sum_{h \neq i} P^h_{ik} \tag{6.23}$$

which he or she equates to his or her own personal MRS between goods k and 1, $\text{MRS}^i_{X_{ik},X_{i1}}$. With the "competitive" side payments received from the $(H - 1)$ other consumers equal to

$$P^h_{ik} = \text{MRS}^h_{X_{ik},X_{h1}}, \quad \text{for all } h \neq i \tag{6.24}$$

the external effect of person i's consumption of good k, and with the producers setting

$$P_k = \text{MRT}_{k,1} \tag{6.25}$$

This expanded competitive market structure satisfies the pareto-optimal condition:

$$\sum_{h=1}^{H} \text{MRS}^h_{X_{ik},X_{h1}} = \text{MRT}_{k,1}$$

Notice that $(H - 1)$ "competitive" side markets (prices) are required just for person i's consumption of good k, plus the usual market between the producer and consumer i, or H markets (prices) in all. By a similar analysis, $H(H - 1)$ additional side markets (prices) would be necessary to allocate X_k correctly among all H consumers, with $(H - 1)$ side markets (prices) for each of the H consumers, a formidable set of markets indeed. Adding the H markets (prices) between the producers and each consumer, there would be H^2 markets (prices) in all. Furthermore, $H(H - 1)$ distinct side markets (prices) and H^2 total markets (prices) are necessary for *each* pure public good.

The same analysis applies if the externality is a diseconomy (a "bad"), although the mutual gains come about indirectly. Achieving pareto optimality increases aggregate real income by moving the economy to the first-best utility−possibilities frontier, and the additional income can then be redistributed to everyone's mutual gain. Each person *i* might be skeptical of this argument, however, and refuse to make "competitive" side payments equal to the marginal damage he or she is causing other people. These payments have the obvious direct effect of lowering the person's utility, and he or she may doubt that he or she will receive adequate compensation when the additional real income is distributed.

Relating the problem of externalities to market failure in this way suggests why the market system breaks down in their presence even when it is mutually beneficial for people to form the necessary side markets. The existence of potential mutual gains from trade is the motivation that normally causes markets to form. For these externalities, however, three difficulties hinder the development of the proper side markets.

The first is that legal and/or political constraints may preclude formation of the side markets, especially in the case of external diseconomies. Assume that an industry in New York State is polluting Vermont air. Even if New Yorkers are convinced that side payments to Vermonters can increase the combined welfare of both states, they certainly have no guarantee that sufficient tax revenues (and other necessary income) will be transferred back to New York to make the potential gain for New Yorkers a real one. Without proper redistributions, New Yorkers may well be better off continuing to pollute, especially if most of the costs of pollution are borne by Vermonters because of prevailing westerly winds. These same circumstances pose difficulties for government intervention in the form of corrective taxes, the textbook solution to externalities to be described in the next section. Who will levy these taxes? Certainly not New York State, and Vermont cannot tax New York citizens for pollution damage. Moreover, the Constitution of the United States may well proscribe levying federal taxes on New York citizens based on damage caused to Vermont citizens. As will be argued in more detail in Part V, a federalist system of governments causes problems for any policy designed to correct for externalities when the externalities spill over jurisdictional lines.

Transactions costs are a second potential hindrance for developing side markets, especially when the external effects are extensive. Suppose people benefit from other people's education but differentially depending on just who is educated (for example, bright people versus dull people). Furthermore, suppose all other people do not benefit equally from any one person's education. In short, education may have the properties of a (virtually) pure public exclusive good, with many different values to the marginal rates of substitution that comprise the external effects. If so, then the sheer number of side markets required to achieve an optimal amount of education for anyone is staggering (H is a very large number) and the costs of even trying to get everyone together are clearly prohibitive, meaning that they would almost surely offset any efficiency gains from achieving or even approaching pareto optimality. Put another way, normal competitive markets permit all consumers and producers to face the same price, an enormous advantage in terms of information requirements. In contrast, externalities of the type under consideration generally require negotiations and differential prices among all consumers in the market, a huge increase in structural complexity. Small wonder, then, that such side markets almost never form, even when the mutual gains, ignoring transactions costs, are obvious to all, such as for external economies. Unfortunately, all important examples of externalities associated with private activity, such as education, pollution, and research and development, affect a very large number of people. After all, their broad scope is what makes them important.

Finally, mutually beneficial side payments might not obtain even if the externality were relatively simple, affecting only a few people, and none of the problems mentioned above existed. There remains the problem that the affected parties have an incentive not to reveal their true preferences. Suppose person *i*'s consumption of good *k* generates an external economy for persons *j* and *m*. Despite the benefits he or she receives, person *j* might decide not to subsidize *i*'s consumption, hoping instead that the other person, *m*, will do so. In the parlance of the literature, *j* desires to be a "free rider." Person *m* reasons similarly, and because no one wants to play the sucker, no side payments occur, despite the obvious gains to all. Various tax schemas exist for avoiding the free-rider problem, but we will defer discussion of them until the next section on nonexclusive goods since the revelation problem has been most closely associated with these goods. It could just as easily apply to exclusive goods, however.

Bargaining and the Coase Theorem

Ronald Coase felt differently about the possibilities for side payments, at least when the externalities involve a small number of consumers or firms. He argued that the appropriate side bargains would take place so long as the property rights to the external effects were established (for instance, someone held the rights to the benefits from a research and development project). His reasoning was simply that bargaining to achieve the pareto-optimal conditions represents a pareto-superior move, and rational, utility-maximizing consumers can be counted on to realize the mutual gains, by the definition of rationality. Some of

the gains may have to be redistributed among the parties to ensure that everyone is better off, but this too is in everyone's mutual interest. Also, the bargained solution does not necessarily have to set each price equal to the MRS, as the competitive market analog suggests. All it must do is select the levels of private activity that satisfies the pareto-optimal conditions, Eqn (6.20), and possibly redistribute some of the gains to ensure that everyone is better off relative to the status quo. Coase's argument became universally known as the Coase Theorem.[6]

The Coase Theorem was a provocative challenge to received public sector theory at the time, which stressed incentives to free ride and presumed that government intervention would always be necessary to achieve pareto optimality in the presence of externalities. The theorem has generated a huge literature, sometimes favorable, sometimes critical. The most recent literature has concentrated on the validity of the Coase Theorem when people have private information about the external effects, and the results have generally been unfavorable to the theorem. This is so even when the external effects are extremely limited, such as to one or two "third parties." Private information is a second-best problem, however, so we will defer most discussion of the Coase Theorem until Chapter 20 in Part III. For now it is enough to note that the Coase Theorem was never assumed to apply when the external effects were extensive. (We will return to it briefly in Chapter 7.)

The Tax/Subsidy Solution

Society does not have to rely on private bargaining to correct for externalities. The government has the option of taxing (subsidizing) externality-generating activities to achieve the pareto-optimal conditions. The tax (subsidy) scheme is simpler than the required competitive market structure, by a factor of H. To see why, consider again the decision by person i to consume (supply) good (factor) k, X_{ik}. As before, assume that all markets are competitive in line with first-best analysis, and that $P_1 = 1$, the numeraire. Person i's decision to purchase good k affects all other people, but for these people it is essentially a lump-sum event. Only person i decides the quantity; the others must accept it as a parameter. Thus, the government needs only to adjust person i's behavior with respect to X_{ik}, as follows.

Before government intervention, all producers and consumers face the same price P_k, which producers set equal to the MRT (marginal cost) and consumers to their personal-use MRS. The government does not want this, but it knows that if it establishes another set of prices for person i (say, P_k^i), then consumer i will set this price to $\text{MRS}^i_{X_{ik},X_{i1}}$, his or her own personal-use MRS. The goal, then, is to design a tax for person i, t_k^i that simultaneously:

1. Drives a wedge between P_k^i and P_k such that:
$$P_k^i = P_k + t_k^i$$
2. Achieves the desired pareto-optimal condition,
$$\sum_{h=1}^{H} \text{MRS}^h_{X_{ik},X_{h1}} = \text{MRT}_{k,1}$$

The proper tax is $t_k^i = -\sum_{h\neq i}\text{MRS}^h_{X_{ik},X_{h1}}$, equal to the sum of the marginal effects on all others of person i's consumption (supply) of good (factor) k. With this tax, the price person i pays for good k, P_k^i, differs from the producer's marginal cost price of k by exactly the summation of his or her marginal effects on all other consumers:

$$P_k^i = \text{MRS}^i_{X_{ik},X_{i1}} = P_k + t_k^i = \text{MRT} - \sum_{h\neq i}\text{MRS}^h_{X_{ik},X_{h1}} \tag{6.26}$$

Thus, the tax establishes the correct pareto-optimal condition on X_{ik}. It is referred to as the Pigovian tax after the British economist A. C. Pigou, who first proposed taxes (subsidies) equal to the sum of the external marginal effects to correct for externalities (Pigou, 1932).

Using the convention that the MRS between two goods is positive if these side effects are beneficial (an external economy), then the "tax" t_k^i is negative, a subsidy, so that person i pays less than the marginal cost price of producing good k. Conversely, if the side effects are harmful (external diseconomies), the tax is positive and person i pays more than the marginal cost price of producing good k.

Furthermore, the government can adjust the tax to the desired level, at least in principle. Suppose at first the tax is zero, and X_{ik} generates external economies. Without benefit of the subsidy, person i consumes (supplies) too little of good (factor) k, and $t_k^i > -\sum_{h\neq i}\text{MRS}^h_{ik,h1}$. A subsidy to person i lowers P_k^i, increases X_{ik}, and thereby decreases the absolute value of each other person's $\text{MRS}^h_{X_{ik},X_{h1}}$. Thus, it is possible to find the t_k^i, such that $t_k^i = -\sum_{h\neq i}\text{MRS}^h_{ik,h1}$ as required.

To allocate the aggregate amount of X_k correctly requires H separate taxes, t_k^i, $i = 1, \ldots, H$, one for each consumer (supplier) of X_k, determined exactly as above. pareto optimality requires

$$\sum_{h=1}^{H} \text{MRS}^h_{X_{ik},X_{h1}} = \text{MRT}_{k,1} \quad \text{all } i = 1, \ldots, H$$

The effects of the tax can also be considered in terms of supply and demand curves (with $P_1 = 1$). Think of k as a good. The aggregate supply curve for good k has the usual interpretation. It is the horizontal summation of the

6. Coase (1960). The assignment of property rights to the activities associated with the external effects is crucial to the theorem. A counterexample is water or air pollution. Private bargaining cannot work here because air and most bodies of water are common-use resources. No one can hold the property rights to clean air and clean water on the public bodies of water.

marginal cost curves for the individual producers of k. The aggregate demand curve for the good k is, similarly, the horizontal summation of the individual consumers' demand curves for good k, with this important difference. Before the individual curves are summed horizontally, they are each adjusted vertically downward (upward) by the amount of the tax (subsidy), t_k^i. Because of the way the taxes (subsidies) are defined, the vertical adjustments just equal $\sum_{h\neq i}\text{MRS}_{ik,h1}^h$ at each unit of X_{ik}, person i's combined marginal impact on all other people. Thus, the resulting individual demand curve for person i reflects the entire $\sum_{h=1}^{H}\text{MRS}_{X_{ik},X_{h1}}^h$, including person i's own MRS between goods k and 1. Because these adjusted curves are then summed *horizontally* to be equated with aggregate supply,

$$\sum_{h=1}^{H}\text{MRS}_{X_{ik},X_{h1}}^h = \text{MRT}_{k,1} = P_k \quad \text{all } i = 1,\ldots,H \tag{6.27}$$

in aggregate equilibrium, as required for pareto optimality. The taxes, t_k^i, if set optimally, determine the effective price for each person i and their individual contribution, X_{ik}, to the aggregate X_k at the equilibrium.

Designing the proper set of H taxes for any one pure public good is obviously a hopeless task. For *each* of the taxes, the government must know $(H-1)$ separate pieces of information, the $\text{MRS}_{ik,h1}^h$. The full set of H taxes, therefore, requires $H(H-1)$ independent pieces of information, all of which may differ. In general, $\text{MRS}_{ik,h1}^h \neq \text{MRS}_{ik,j1}^j$, for $j \neq h$. (Think of the fence example. The external MRS effect on each third party depends on how their view is affected by whoever builds the fence.) Finally, a world consisting of one pure private good and $(N-1)$ pure public goods would require $H(N-1)$ taxes and $H(H-1)(N-1)$ independent observations on the external marginal effects.

Limited Externalities

Only a small subset of people is likely to be affected by the consumption of some good; that is, the good is somewhere on the continuum between pure publicness and pure privateness. (For example, a fence is likely to affect only the neighbors on the adjacent properties and perhaps not all of them.) As is immediately obvious from the construction of the model, the pareto-optimal rule:

$$\sum_{h}\text{MRS}_{ik,h1}^h = \text{MRT}_{k,1}$$

applies only to the subset of H people affected by person i's consumption of good k. The subset could number as few as two people (person i and one other), and pareto optimality would still be described by this rule. Furthermore, the subset of people whose consumption generates consumer externalities could number far fewer than H people. There may only be one such person. As a practical matter, the government would only intervene if the number of people affected by a particular externality was fairly large and/or the externalities generated in any one instance were deemed to be "substantial" in some sense. Very few goods (factors) are likely to meet this practical criterion. That is, most goods are certainly well toward the pole of pure privateness. Private bargaining may be the preferred solution when the numbers affected are small if, indeed, any action can hope to improve the private market outcomes given the transactions costs of bargaining or government intervention.

All these considerations serve to mitigate the actual policy problems caused by consumption externalities. Nonetheless, if, for example, J people were affected by each of L goods as described by the model, then $J\cdot L$ taxes (subsidies) are required for allocative efficiency, a formidable task even if both J and L are "fairly small" relative to all the people and all the goods and factors in the economy.

We have been analyzing the case of *individualized externalities*, in which the external effects associated with private sector activity depend not only on what the activity is but who is doing it: It matters who builds a fence. The inescapable conclusion is that neither government taxes and subsidies nor private bargaining can be expected to achieve the pareto-optimal conditions for any individualized externality in which the external effects are widespread.

Not all externalities are individualized, however. The final two sections of the chapter consider two common types of externalities that are not individualized: the nonexclusive good and the aggregate externality. The aggregate externality is the more hopeful of the two from a policy perspective.

NONEXCLUSIVE GOODS—THE SAMUELSON MODEL

Paul Samuelson was the first economist to analyze the problem of externalities using a formal general equilibrium model of social welfare maximization for his analytical framework. He developed his model in three articles published in the 1950s, Samuelson (1954, 1955, 1958) and it is safe to say that no other single work has been more influential to the development of public expenditure theory. For this reason alone, his model deserves special attention in any treatise on public sector economics. It also happens to be a useful vehicle for exploring a number of important issues, including:

1. The special problems caused by nonexclusive goods, Samuelson chose the nonexclusive good for his example of an externality.
2. A method for introducing the government into the standard general equilibrium model, given that the

government's preferences are not supposed to count other than in providing the social welfare function.

3. The important first-best dichotomy property that a competitive market system correctly allocates pure private goods. This property could have been developed above by considering a model with at least two pure private goods. It always holds under first-best assumptions.
4. An initial presentation of the *benefits-received* principle of taxation, one of the two widely accepted normative criteria for judging whether or not a particular tax is fair.

A nonexclusive good (a service, really) has the property that if any one person consumes it, everyone necessarily consumes its services in equal amounts. Nonexclusivity works both ways. On the one hand, if one person consumes the good, he or she cannot exclude others from consuming it. On the other hand, once someone consumes the good, no individual within the domain of the good can exclude himself or herself from consuming the services of the good even if he or she should want to. Consumption is truly joint. These goods cause terrible problems for any society dedicated to competitive market principles and consumer sovereignty. Unfortunately, they are hardly theoretical curiosities to be found only in obscure economics journals. Defense, the exploration of outer space, and global warming are three very important examples of nonexclusive goods.[7]

The free-rider problem undermines the ability of markets to allocate nonexclusive goods. Markets work for exclusive goods because people must purchase the goods to receive any utility from them. They reveal their preferences when they purchase the goods. In contrast, people do not have to purchase a nonexclusive good to receive its services. The strategy of free riding is a viable, and preferred, option. People have an incentive not to reveal their preferences, hoping that someone who wants the good will actually buy it. If someone does play the "sucker," everyone immediately consumes its services as free riders. Therefore the government is forced to purchase the good on behalf of society for there to be any hope of achieving the proper allocation of resources to the good, and perhaps to have any of the good at all, even though everyone might desire the services of the good. This is why Samuelson labeled nonexclusive goods as "public goods." As we shall see, these goods satisfy our definition of public goods, which can also apply to exclusive goods. Samuelson's equation of "publicness" with "nonexclusivity" (joint consumption) is the most often employed one in the externality literature, however.

Having decided to purchase the good, the government is faced with two difficult questions:

1. How much of the good should it buy?
2. How should people be taxed to pay for the good?

One can provide answers to both questions consistent with the standard criteria of consumer sovereignty, pareto optimality, and competitive market principles, but these answers depend upon consumers revealing their true preferences to the government. Unfortunately, consumers have no more incentive to relate their true preferences to the government than they do to the marketplace. In answering these questions, therefore, the government confronts the mechanism design problem. It must find a tax scheme that induces consumers to reveal their preferences.

The Government in a General Equilibrium Model

To focus on the problems peculiar to nonexclusivity, assume that there is one nonexclusive good, the kth, and that all other $(N - 1)$ goods are pure private goods. Assume further that the market for nonexclusive goods is inoperative because of the free-rider problem, so that the government must decide how much of the good to buy and how to ask people to pay for it. The immediate problem, then, is to incorporate the government into the formal model of social welfare maximization.

One method of proceeding is to assume that the government has a preference function for nonexclusive goods derived through some sort of political process, exactly the approach taken for the government's social welfare function. If this government preference function also includes the overall size of the private sector as one of its arguments, then the private sector defines the opportunity costs of public expenditures on the nonexclusive good, and finding the optimal amount of the "public" good becomes a simple exercise in consumer theory. The government would solve a problem of the general form:

$$\max_{G(\text{public expenditures},\text{private sector})} \left(\text{public expenditures}, \text{private sector}\right)$$
$$\text{s.t. Public expenditures} + \text{Private sector} = Y$$

where G = the government's preference function and Y = total national product to be split among the private and public sectors.

As was stressed in Chapter 1, however, the government is not supposed to interject its own preferences into the decision-making process according to the mainstream normative theory. Rather, it is supposed to play the part of

7. The terminology "nonexclusive" introduced by Samuelson is somewhat misleading as some good might have the properties described above over a subset of individuals yet not be available at all to still other people. Compare national defense with local police protection. Jointness of consumption is perhaps a more accurate description, leaving open the possibility that some consumers may be excluded. At this point, however, we will use *nonexclusiveness* and *jointness* in consumption interchangeably and assume that the entire population is affected.

agent, acting solely upon consumers' preferences for its demand data whenever possible—that is, to honor the principle of consumer sovereignty.

The government has no choice but to violate consumer sovereignty when faced with the distribution question. Society must develop a set of social welfare rankings through some political processes that establish the criteria for achieving end-results equity. Individual preferences, by themselves, are not sufficient to determine the interpersonal equity conditions for the optimal distribution of income. But such is not the case with allocational issues. Consumer preferences are sufficient to determine the demand component of the pareto-optimal conditions for allocational efficiency, and consumers have preferences over all goods and services, including nonexclusive goods. There is no reason why their preferences cannot be honored, at least in principle. Thus, mainstream normative theory has rejected the construct of a distinct government preference function for nonexclusive goods, or, indeed, for any expenditures arising for allocational reasons. Only consumers' preferences enter the optimal normative policy rules.

A simple analytical device for introducing government purchases into the standard general equilibrium model without generating a distinct government demand for these purchases is to define a fictitious individual (say, the first) to represent the government.[8] By fictitious we mean that $U^1(\vec{\mathbf{X}}_1) \equiv 0$, where $\vec{\mathbf{X}}_1$ is a vector of government purchases. The vector $\vec{\mathbf{X}}_1$ enters into the production–possibility frontier and market clearance—the goods themselves are real and use up scarce resources—but government preferences never count for social welfare, as $\partial U^1/\partial X_{1k} = 0$, for any k.

Allocating a Nonexclusive Good

If good k is the only nonexclusive good, social welfare maximization can be represented as

$$\max_{(X_{1k};X_{hj})} W\left[U^1(X_{1k}); U^h(X_{hj}; X_{1k})\right]$$
$$\text{s.t. } F\left(\sum_{h=2}^{H} X_{hj}, X_{1k}\right) = 0$$

where

X_{1k} = the nonexclusive good, purchased only by the government.
X_{hj} = good (factor) j consumed (supplied) by person h, $h = 2, \ldots, H_j$; $j = 1, \ldots, k-1, k+1, \ldots, N$.
U^1 = the (fictitious) preference function of the government.

8. This technique was first demonstrated to us by Peter Diamond in a set of unpublished class notes.

The corresponding Lagrangian equation is

$$\max_{(X_{1k};X_{hj})} L = W\left[U^1(X_{1k}); U^h(X_{hj}; X_{1k})\right] + \lambda F\left(\sum_{h=2}^{H} X_{hj}, X_{1k}\right)$$

Notice that even though only the government purchases good k, X_{1k} enters into each person's utility function since everyone automatically consumes the entire services of the nonexclusive good. Compare this with the case of an exclusive good that generates externalities. With the exclusive good, there is a distinct difference between the services it provides privately to each individual who purchases it and the flow of external services received by other consumers in the form of external economies or diseconomies. Think once again of the fence. The person who built the fence receives a flow of services that are distinct from the "services" bestowed upon his or her neighbors. With nonexclusive goods, however, there is no such distinction. Whatever services are available to the purchaser, these identical services are automatically available to all others, whether they want the services or not.

Interpersonal Equity Conditions

Consider the necessary conditions for a social welfare maximum for the model with nonexclusive goods, beginning with the interpersonal equity conditions for good 1. Take the first-order conditions with respect to X_{h1} and X_{21}, the consumption of good 1 by persons h and 2:

$$X_{h1}: \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{h1}} = -\lambda F_1 \tag{6.28}$$

$$X_{21}: \frac{\partial W}{\partial U^2}\frac{\partial U^2}{\partial X_{21}} = -\lambda F_1 \tag{6.29}$$

Consequently,

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{h1}} = -\lambda F_1 \quad \text{all } h = 2, \ldots, H \tag{6.30}$$

Equation (6.30) is the standard result that good 1 should be distributed lump sum across all individuals to equalize the social marginal utility of good 1.

Pareto-Optimal Conditions

The pareto-optimal conditions are obtained somewhat differently in this model. We have to compare the government's purchase of good k, X_{1k}, with any other consumer's purchase of any private good—say, the purchase of good 1 by person j, X_{j1}. The first-order conditions are

$$X_{1k}: \sum_{h=2}^{H} \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{1k}} = -\lambda F_k \tag{6.31}$$

(Recall that $\partial U^1/\partial X_{1k} = 0$)

$$X_{j1}: \frac{\partial W}{\partial U^j}\ \frac{\partial U^j}{\partial X_{j1}} = -\lambda F_1 \tag{6.32}$$

Dividing Eqn (6.31) by Eqn (6.32),

$$\frac{\sum_{h=2}^{H} \frac{\partial W}{\partial U^h}\ \frac{\partial U^h}{\partial X_{1k}}}{\frac{\partial W}{\partial U^j}\ \frac{\partial U^j}{\partial X_{j1}}} = \frac{F_k}{F_1} = \mathrm{MRT}_{X_k,X_1} \tag{6.33}$$

But, if the government has correctly redistributed good 1 such that the interpersonal equity conditions hold, then

$$\frac{\partial W}{\partial U^j}\ \frac{\partial U^j}{\partial X_{j1}} = \quad \text{all } j = 2,\ldots,H$$

Selectively substituting for the denominators in each term on the LHS of Eqn (6.33) and canceling $\partial W/\partial U^h$ term by term yields

$$\sum_{h=2}^{H}\left(\frac{\frac{\partial U^h}{\partial X_{1k}}}{\frac{\partial U^h}{\partial X_{h1}}}\right) = \mathrm{MRT}_{X_k,X_1} \tag{6.34}$$

or

$$\sum_{h=2}^{H}\mathrm{MRS}^h_{X_{1k},X_{h1}} = \mathrm{MRT}_{X_k,X_1} \tag{6.35}$$

Condition (6.35) gives the familiar result that the sum of each person's MRTs between the nonexclusive good and good 1 equals the MRT between X_k and X_1 in production. Samuelson was the first to demonstrate formally the summation rule for externalities. Therefore, Eqn (6.35) is commonly referred to as the Samuelson Rule. Subsequent research showed that this same type of rule also applies to exclusive goods that generate externalities, as already demonstrated above for individualized externalities.

The First-Best Dichotomy: The Private Goods and Factors

Before discussing the government's prospects of satisfying the pareto-optimal conditions, consider the following important proposition: In a first-best policy environment, pure private goods and factors can be allocated efficiently by the competitive market system despite the presence of externalities elsewhere in the economy. To see this, consider the pareto-optimal conditions for any two pure private goods (or factors). Compare, for example, the first-order conditions for X_{hm} and X_{h1}, two private goods (factors) consumed (supplied) by person h. The first-order conditions are

$$X_{hm}: \frac{\partial W}{\partial U^h}\ \frac{\partial U^h}{\partial X_{hm}} = -\lambda F_m \tag{6.36}$$

$$X_{h1}: \frac{\partial W}{\partial U^h}\ \frac{\partial U^h}{\partial X_{h1}} = -\lambda F_1 \tag{6.37}$$

Dividing Eqn (6.36) by Eqn (6.37),

$$\frac{\frac{\partial U^h}{\partial X_{hm}}}{\frac{\partial U^h}{\partial X_{h1}}} = \frac{F_m}{F_1} \quad \text{all } h = 2,\ldots,H \quad \text{any } m \neq k \tag{6.38}$$

or

$$\mathrm{MRS}^h_{X_m,X_1} = \mathrm{MRT}_{X_m,X_1} \tag{6.39}$$

These are the standard pareto-optimal conditions for private goods developed in Chapter 2 and they are achieved by competitive markets for m and 1. Therefore, the existence of nonexclusive goods does not upset the competitive allocations of the other pure private goods, at least with first-best assumptions. That this property applies to our model of externality-generating exclusive goods should be clear from the structural similarities of the two models.

Policy Problems with Nonexclusive Goods

Knowing that it should purchase a nonexclusive good to the point at which $\sum_{h=2}^{H}\mathrm{MRS}^h_{X_{1k},X_{h1}} = \mathrm{MRT}_{X_k,X_1}$ may not be very helpful for the government in practice, as it still has the vexing problem of determining each person's MRS under the handicap of nonrevelation. The problem is not that an MRS is a special theoretical construct that cannot be observed in practice. For pure private goods, its value is easily determined. Assuming rational behavior, the MRS between any two goods for any consumer simply equals the price ratio of the two goods. Rather, the problem is nonexclusivity itself, which leads to the incentive to free ride. Competitive market analogs to private goods are of little help to the government.

The government cannot simply set a price (tax), ask consumers how much they would be willing to buy at that price, and compare quantities demanded with quantities supplied at the producer price to check for equilibrium. Consumers might well hide their preferences if they thought they might actually have to buy the stated amounts at the going price. Furthermore, this competitive process would not generate the pareto-optimal quantity even if revelation were not a problem. The market process is reversed for nonexclusive goods: the single output selected by the government is the given for each consumer, not the price. Therefore, the proper method of reaching equilibrium is for the government to select an *output*, ask consumers how much they would be willing to pay for the last unit of the output, add each consumer price, and compare the aggregate consumer demand price with the marginal cost (the producer's supply price) at the selected output. The

optimum quantity occurs at the output for which the aggregate demand price equals the supply price.

In terms of the standard supply and demand diagram, every consumer has a demand curve for the nonexclusive good even if he or she would not reveal it. Just as with the externality portion of exclusive goods, these demand curves must be added *vertically*, not horizontally, to arrive at aggregate market demand. (There is no further horizontal summation, however, as the quantity selected by the government *is* the aggregate quantity.) The quantity at which the vertical summation of individual demand curves intersects the supply curve satisfies the pareto-optimal condition, Eqn (6.35).

This reversed competitive process is illustrated in Fig. 6.2 for the two-person case. In the diagram:

1. d_k^1 and d_k^2 are the individual's demand curves for X_k, reflecting their respective $\text{MRS}_{k,il}^i$ at every X_k ($P_1 = 1$, the numeraire).
2. D_k^{total} is the vertical summation of d_k^1 and d_k^2.
3. S_k is a normal supply curve for X_k reflecting the $\text{MRT}_{k,1}(\text{MC}_k)$ at every X_k.
4. P_k is the producer's supply price at X_k^{opt}.
5. $P_k^2 = \text{MRS}_{X_k,X_{21}}^2$ at X_k^{opt}.
6. $P_k^1 = \text{MRS}_{X_k,X_{11}}^1$ at X_k^{opt}.

At X_k^{opt},

$$P_k = P_k^1 + P_k^2 \tag{6.40}$$

or

$$\text{MRT}_{X_k,X_1} = \text{MRS}_{X_k,X_{11}}^1 + \text{MRS}_{X_k,X_{21}}^2 \tag{6.41}$$

Thinking in terms of defense, if the last weapon system costs \$20 billion, and in the aggregate, consumers are willing to pay \$20 billion based on their marginal rates of substitution, then the defense budget is optimal.

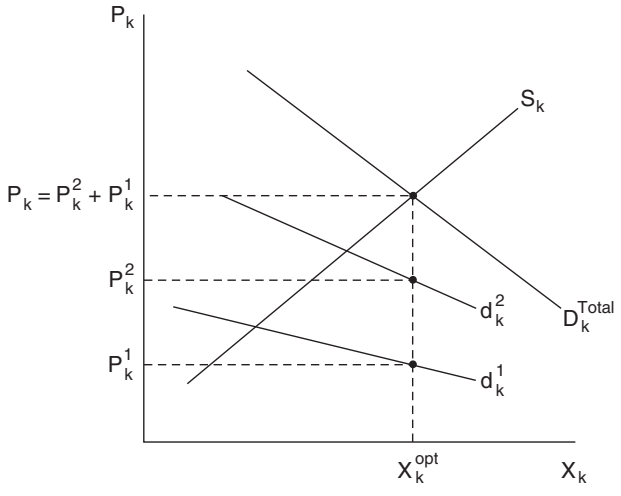

FIGURE 6.2

Thus, it is possible to describe a competitive analog for establishing the optimum quantity of nonexclusive goods, but the analog is not terribly useful in practice as consumers have little incentive to reveal their demand prices at each quantity. The government has to design a different mechanism to induce consumers to reveal their preferences for nonexclusive goods. Otherwise, the government has little choice but to select a quantity and hope that its choice is correct without benefit of the normal market signals to aid its judgment.

Paying for the Public Good

The question "How should people be taxed to pay for a nonexclusive good such as defense?" can be viewed as uninteresting from a normative perspective. It has no normative significance for social welfare maximization in the mainstream perspective. The only normative requirement for the government is to select the optimal quantity of the good. Suppose it has. The government's output choice is exogenous from each consumer's point of view and, since the government is not interfering in any other market, the pareto-optimal conditions for all other goods (factors) hold as well. Therefore, all the government need do to preserve efficiency is raise taxes on a lump-sum basis to finance the good. Any lump-sum tax will do—for example, a head tax based on age scaled up or down until sufficient revenues have been collected.

The only caveat is that the optimal quantity of the nonexclusive good depends in part on the particular lump-sum tax chosen. Any tax shifts people's demands for the nonexclusive good simply because their incomes have changed. The new pattern of after-tax incomes therefore dictates a new output choice to satisfy the pareto-optimal condition for the nonexclusive good. But any pattern of lump-sum taxes allows all the pareto-optimal conditions to hold, by the definition of a lump-sum tax. The "only" allocational problem for the government remains selecting the correct output for the chosen tax. Furthermore, any adverse distributional consequences of a particular tax such as an age tax would be fully offset by the lump-sum redistributions that satisfy the interpersonal equity conditions. In this sense, then, the question of how people should pay for the good is uninteresting; it can be entirely subsumed within the distribution question.

Public sector economists have nonetheless expressed considerable interest in the payments mechanism, for equity and efficiency reasons. The equity motivation is that citizens may not accept any pattern of lump-sum taxation to pay for these goods, especially if no strong consensus has emerged regarding the social welfare function. They may well insist that the taxes satisfy commonly held notions of equity, that they be fair as well as efficient. The efficiency motivation is the mechanism design problem. Finding a tax

scheme that induces people to reveal their true preferences for the nonexclusive good is essential. Avoiding the free-rider problem removes the principal barrier toward achieving the pareto-optimal allocation. Should these taxes also be deemed equitable, so much the better. Let us consider the question of equity first.

The Benefits-Received Principle of Taxation

Although there are no equity norms agreed upon by everyone, two general principles of fair taxation have gained remarkably wide acceptance in Western economic thought as practical guidelines for tax policy. Taxes are deemed fair if they are related to the benefits received from public goods and services, or if they are closely related to each person's ability to pay.

The benefits-received principle of taxation is the older of the two principles. It dates back at least to the fourteenth and fifteenth centuries in European feudal societies, when the nobles paid a tribute to the king in return for protection from foreign enemies. The benefits-received principle is meant to apply to all resource-using public expenditures, such as nonexclusive goods. It is especially compelling in capitalist societies as a natural and fair way to pay for public services because the payment for goods in the marketplace is on a benefit-received basis. The rationale for taxing, according to the benefits received from public services, runs as follows:

> The government is engaged in allocational activities only because one of the technical assumptions underlying a well-functioning market system fails to hold and the competitive market system is signaling an incorrect allocation of resources. Because the government is merely substituting for the competitive market system in these instances, taxes raised to finance these activities should imitate the quid pro quo feature of market prices. Competitive markets exact payments from consumers and producers reflecting the benefits received from their market transactions. Thus, taxes should reflect the benefits received from the government services.

The benefits-received principle is obviously limited to the allocational, or resource-using, part of the government's budget. Transfer payments designed to achieve distributional goals cannot possibly be financed by the benefits principle because the transfer recipients are the primary beneficiaries of the transfers. Consequently, public sector economists have developed a second practical guideline for equitable taxation, the ability-to-pay criterion, first proposed by Adam Smith and John Stuart Mill in the late 1700s and the early 1800s. Smith (1904) and Mill (1921) viewed taxes as a necessary sacrifice that citizens undertake to support the commonwealth, the common good. In their view, people should be asked to sacrifice in accordance with their ability to pay. Their ability-to-pay principle was meant to apply to transfer programs and also serve as the default option for allocational expenditures whenever taxes cannot easily be related to benefits received.

The ability-to-pay principle is clearly related to society's distributional norms and bears a kinship to the modern social welfare view of distributive justice. We will discuss it in detail in Chapter 11. Our present goal is to consider tax schemes designed to finance expenditures on nonexclusive goods, for which the benefits-received principle is meant to apply.

Saying that taxes should be related to the benefits received from public expenditures is still too general for policy purposes. It begs the immediate question of exactly what benefits should be used as the basis for taxation: total benefits?, average benefits?, marginal benefits?, and so forth. There is less agreement on this question than on the general principle itself, but one can make an excellent case for choosing marginal benefits as the appropriate tax base. If society firmly believes in competitive market principles and views the government as an agent merely substituting for the market in any of these allocational areas, then a tax system that duplicates competitive pricing principles is likely to be considered fair by that society. Competitive prices equal marginal benefits, more accurately consumers' (producers') marginal rates of substitution (marginal rates of transformation) between any two goods (factors). Therefore, taxes that equal marginal rates of substitution are truly pseudo-competitive prices. Whether one labels them taxes or prices hardly matters.

Following this competitive interpretation of the benefits-received principle, the government ideally should levy a set of H differential taxes to pay for the nonexclusive good, equal to each person's MRS between the good and a private (numeraire) good at the quantity selected by the government. In terms of Fig. 6.2, person 2 would pay a tax $t_k^2 = P_k^2$, and person 1 a tax $t_k^1 = P_k^1$. At the optimum, these taxes would add exactly to the supply price P_k, equal to the marginal cost of producing X_k. Taxing or pricing in this way is known as Lindahl pricing after the Swedish economist Eric Lindahl, who first proposed this method of taxation.[9] Lindahl prices have the dual properties of preserving allocational efficiency *and* satisfying widely held notions of tax equity because of their direct correspondence with competitive market pricing.

Notice the kinship between Lindahl prices and Pigovian taxes levied on externality-generating exclusive goods. Pigovian taxes are also benefits-received taxes in the sense that they equal the aggregate marginal external benefit (damage) resulting from the consumption of the exclusive good. These

9. Lindahl (1958). See also subsequent developments in Johansen (1963), Samuelson (1969).

taxes (subsidies) *have* to be equal to the aggregate marginal damage (benefit) to achieve pareto optimality.

Interpreting the benefits received as marginal benefits received is required for most allocation problems, as one would suspect. Nonexclusive goods happen to be an exception, however. We have seen that Lindahl prices are not necessary for achieving pareto optimality with nonexclusive goods; any lump-sum tax also supports the optimum. But Lindahl prices do support the efficient allocation and their basic appeal is one of equity, that they represent a competitive interpretation of the benefits-received principle of taxation.

One might ask how Lindahl prices can be said to imitate competitive pricing, since everyone faces the same price in the market system, whereas Lindahl prices generally differ for each person. The answer lies in the peculiar way in which nonexclusive goods must be marketed, described above. For exclusive goods, price is the parameter faced by all consumers in common. Each person buys the quantity for which price equals the MRS (with the numeraire good as the basis of comparison). Hence, in equilibrium, marginal rates of substitution are equal for all consumers, but the quantities purchased generally differ. The situation is reversed for the nonexclusive good. Everyone is forced to consume the one quantity selected by the government. Because people's tastes differ, their marginal rates of substitution generally differ at that quantity, implying that the price (tax) each should pay differs as well. The competitive pricing principle common to both goods is that price equals the MRS, each person's willingness to pay on the margin.

Virtually any pattern of differential taxes is consistent with competitive pricing applied to nonexclusive goods, since the prices depend only on the individual demands for the good. Return to the two-person example of Fig. 6.2. Person 1 may well place a value of zero on the marginal unit at the optimal quantity, as pictured in Fig. 6.3. If the quantity X_k^{opt} is pareto optimal, then $t_k^2 = P_k$ as drawn. Person 2 would pay the entire cost of the good, and person 1 would pay nothing, even though in a total or average sense, he or she benefits from having the good, as evidenced by his or her willingness to pay positive prices for inframarginal units of the good. In fact, depending on the slopes of d_k^1 and d_k^2, person 1 may actually be willing to pay more for X_k on an average, per-unit basis than person 2, even though his or her marginal evaluation of the good is zero. Thus, a tax schema based on marginal benefits can produce completely different results from one based on total or average benefits received.

It could also happen that, at the optimum, person 1 believes that the government has purchased too much X_k; the marginal units are harmful in his or her view. If this were true, then $t_k^1 < 0$ and $t_k^2 > P_k$, as shown in Fig. 6.4. Person 2 pays *more* per unit than competitive producers require to supply the good, and subsidizes person 1 for the harm caused him or her on the margin at the equilibrium. Notice that the subsidy has nothing whatsoever to do with standard distributional issues. It simply reflects taxes set equal to marginal rates of substitution.

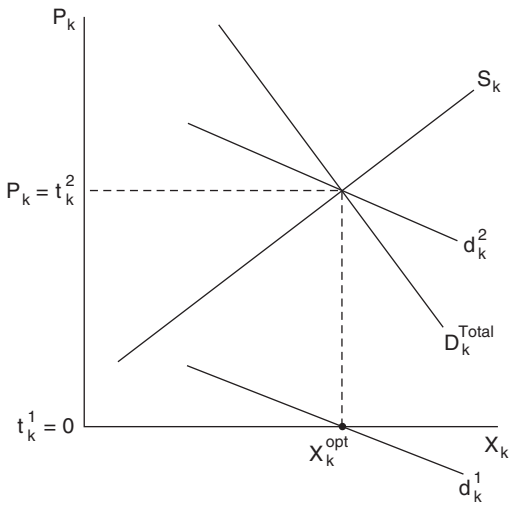

FIGURE 6.3

This situation is hardly an anomaly—it almost certainly applies to defense spending, at least in the United States. Some people believe that the defense budget is much too large and causes them harm on the margin. Others just as clearly believe that the defense budget is too low. They would accept an increase in their current tax burdens if they could be assured that the taxes would be spent on defense.

In the late 1960s, some people refused to pay their federal income taxes in protest against the war in Vietnam. Their protest highlighted one of the problems peculiar to

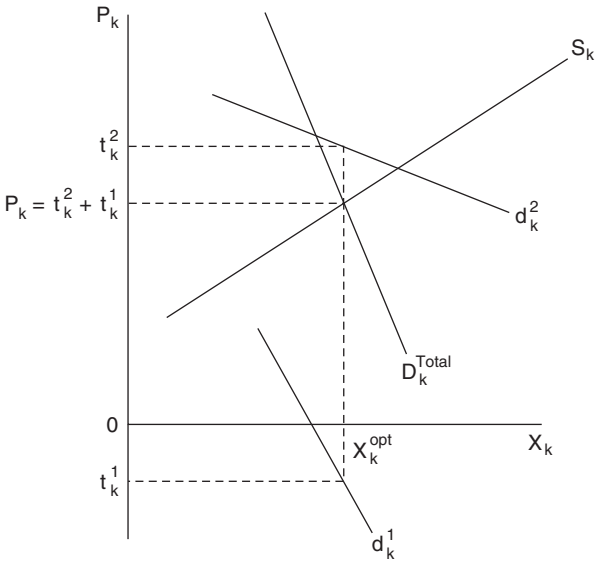

FIGURE 6.4

nonexclusive goods. If consumers do not want an exclusive good, they simply choose not to buy it. This choice does not exist for the nonexclusive goods, but at the very least these protesters felt entirely justified in not paying to support the U.S. effort in Vietnam. On the whole, these people were probably not staunch believers in competitive market principles, yet these principles supported their protest rather well. One wonders how much of a subsidy would have been required to offset the harm done to them, and whether the war effort really was pareto optimal. Some of these people obviously had extremely negative marginal rates of substitution. In principle, even a relatively few negative marginal rates of substitution could generate an $X_k^{\text{opt}} \approx 0$ according to the pareto-optimal rule, if they were extremely negative. The war protest turned on other ethical issues; the protesters did not use the principles of competitive market pricing to support their cause, but they could have.

At the same time, many who supported an even stronger U.S. military effort in Vietnam undoubtedly believed very strongly in competitive pricing principles. Would they have been willing to pay a Lindahl subsidy to the protesters consistent with these principles? Probably not, the point being that the commitment to a competitive market interpretation of the benefits-received principle may not be very strong, despite its underlying rationale. It can easily be overridden by other ethical principles, such as the principle that everyone ought to support the country in time of U.S. military involvement.

People's commitments to various ethical principles may well become confused even on more narrow economic grounds. For instance, people may simply reject the notion of differential payments for goods commonly consumed. A principle of equal payment for equal consumption may appeal to many people's sense of equity, even though this criterion bears no close relationship to competitive market pricing principles in which marginal benefit, not consumption, is what counts. Moreover an appeal to pure economic theory cannot resolve these confusions. Recall that *any* payment schema for nonexclusive goods is consistent with pareto optimality, so long as it is lump sum. A benefits-received principle consistent with competitive market principles is required in other contexts to promote economic efficiency, but not here.

In conclusion, the discussion of Lindahl pricing as a benefits-received tax points out that, strictly speaking, the benefits-received principle has no standing as an equity principle in the mainstream neoclassical model of social welfare maximization. Its only function is to promote efficiency. *All* end-results equity considerations in the mainstream neoclassical model are contained in the social welfare function and the corresponding interpersonal equity conditions.[10] The social welfare function bears no relationship at all to any benefits-received tax, including Lindahl prices. Nonetheless, benefits received as an equity principle was well established in the public sector literature before Samuelson formalized the neoclassical model, and it undoubtedly retains its appeal among the general public as a fair method of taxation.

10. We are grateful to Robin Broadway for emphasizing this point.

Preference Revelation and Taxation: The Mechanism Design Problem

In 1971, Edward Clarke achieved a significant theoretical breakthrough by describing a set of taxes that would, in principle, avoid the free-rider problem with nonexclusive goods.[11] His schema of necessity breaks with the competitive pricing model, which, as we have seen, offers no incentive for people to reveal their preferences. Rather, his taxes are based on the premise that individuals will reveal their true preferences, if they are forced to accept the consequences of their actions on everyone else. The so-called *Clarke taxes* are designed as follows.

Assume that the nonexclusive good, X_k, is competitively supplied at constant cost, with $P_k = \text{MC}_k$. Without loss of generality, set $P_k \equiv \$1$. The government begins by assigning arbitrary per-unit tax shares t_h to each person, with $\sum_{h=1}^{H} t_h = 1$. It then asks everyone to announce their demand curves for the public good, d_k^h. Ordinarily, the intersection of the horizontal price line, \$1, and the vertical summation of the individual, d_k^h, would determine the optimal quantity of the public good, but the government has no reason to believe that the consumers have revealed their true demand curves. This is where Clarke's tax scheme comes into play. It is a mechanism for extracting the individuals' true preferences one person at a time.

Suppose the government begins with person i, all announced demand curves except person i's are summed vertically, and the government selects the quantity given by the intersection of this new aggregate demand curve and $(\$1 - t_i)$, the combined tax share of the other $(H - 1)$ individuals (refer to Fig. 6.5):

AD = vertical summation of all H announced demand curves.
AD $- d_i$ = vertical summation of all but person i's announced demand curve.
t_i = assigned tax share of person i.
$\$1 - t_i$ = combined assigned tax shares of the other $(H - 1)$ individuals.

11. Clarke (1971), Clarke (1972). The discussion in the text closely follows the presentation of "Clarke taxes" by Nicolaus Tideman and Gordon Tullock in Tideman and Tullock (1976). By now, a number of preference-revelation mechanisms have appeared in the literature. For an alternative tax schema applicable to many public goods simultaneously, see Groves and Loeb (1975).

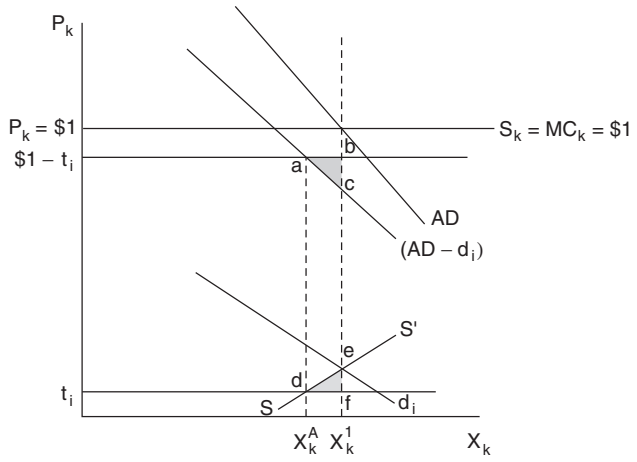

FIGURE 6.5

X_k^A is the initial quantity chosen at the intersection of (AD $-$ d_i) and ($1 $-$ t_i).

Person i is then given the following choice. He or she can accept X_k^A with a total tax payment, $t_i X_k^A$. Alternatively, the government will increase (decrease) the amount of X_k, providing person i pays an additional Clarke tax (receives an additional Clarke subsidy) equal to the amount required to make all the other individuals indifferent to the change, given their announced demands. For instance, should person i vote an increase to X_k^1, his or her Clarke tax would be equal to the triangle abc. Triangle abc is the difference between the total taxes paid by all the other people for the increment $(X_k^1 - X_k^A)$, less the total value of the increment to them as measured by the area under (AD $-$ d_i) between X_k^1 and X_k^A.[12] Draw SS′ through X_k^A as the mirror image of (AD $-$ d_i), so that the area between ($1 $-$ t_i) and (AD $-$ d_i) equals the area between SS′ and t_i. Using SS′, person i's Clarke tax (subsidy) equals the area between SS′ and t_i, area def in the example.

Person i always chooses to reveal his or her true preferences and pay the Clarke tax (subsidy) (unless d_i is horizontal at t_i). As drawn in Fig. 6.5, if d_i is his or her true demand curve, person i chooses X_k^1 and pays the Clarke tax def, in addition to the assigned tax share, $t_i X_k^1$. The marginal benefits and marginal costs of X_k to person i are equal at X_k^1. Any other choice reduces the net benefits from consuming X_k. For example, if d_i were false it would benefit person i to select the intersection of the true d_i and SS′ and pay the corresponding Clarke tax (receive the corresponding Clarke subsidy) along with the assigned tax share. Hence, self-interest dictates true revelation of preferences.

12. We are assuming no income effects, in which case the actual and compensated demand curves are the same, and triangle abc is an appropriate measure of loss suffered by all the other individuals.

What holds for person i holds for everyone. Place the true d_i in AD and go on to the next person. Offer X_k^1 or any other output the person wants, subject to paying the Clarke tax. That person has the same incentive as person i to reveal his or her true demand curve and pay the Clarke tax. Continue until all but one person have been given this option, and let Fig. 6.5 represent the situation at this point. The aggregate demand curve (AD $-$ d_i) now contains the true demand curves for the other (H $-$ 1) individuals, not the original announced demand curves. Therefore, when the last individual has chosen, all the true demand curves have been revealed and the intersection of AD and S^k is the pareto optimum.

The Clarke tax schema bears no necessary relationship to Lindahl prices or any other tax schema that might be deemed equitable, because the assignment of initial tax shares is entirely arbitrary. Tideman and Tullock argue, however, that Clarke taxes could be made consistent with Lindahl prices by letting one citizen assign the tax shares under the condition that the assignor pays a penalty equal to some proportion of the aggregate Clarke taxes at the optimum (Tideman and Tullock, 1976, p. 1156). Presumably this person would have an incentive to minimize Clarke taxes, which implies reassigning tax shares as closely as possible to each person's true marginal evaluation. Referring to Fig. 6.5, person i's tax share would be reassigned to the intersection of d_i and SS′. With the Tideman and Tullock modification, then, people reveal their true preferences by means of the Clarke tax schema, the government chooses the pareto-optimal allocation of X_k, and tax payments correspond to Lindahl pricing, the competitive interpretation of the benefits-received principle of taxation.

Clarke's tax schema was a significant breakthrough in the theory of mechanism design, which is concerned with the problem of how to induce people to tell the truth when they have private information. At the same time, his schema is unlikely to have much practical significance. A government could hardly be expected to administer the Clarke taxes even approximately over a large population; the computational requirements are enormous. And even if it could, Tideman and Tullock note that each individual Clarke tax is likely to be quite small, enough so that many people might actually abstain from voting for a new allocation. They also show how coalitions might form to undermine its revelation properties (Tideman and Tullock, 1976, pp. 1156–1158). Finally, the Clarke tax schema ignores income effects. One must conclude that Clarke taxes do not resolve the free-rider problem as a practical matter.

Do People Free Ride?

The question remains whether people do attempt to free ride on the goodwill of others when they have an

incentive to do so. The mainstream normative public sector theory would dearly prefer that they do not. Truth telling and cooperation in the name of good citizenship are fundamental to the mainstream theory. The notion of the government acting as an agent to promote efficiency when markets fail requires that people tell the truth about their preferences. And the social welfare function, which is so central to the mainstream theory, also presumes cooperative, other-directed behavior when people enter the political sphere to confront the problem of distributive justice.

A number of economists have explored the extent of free riding with nonexclusive goods in experimental settings. They often choose undergraduate economics majors as their test subjects, presumably because economics majors ought to understand the personal advantages of free riding. The results of these experiments are somewhat encouraging to the mainstream theory.

The standard experiment consists of a group of N players who are each given a fixed number of tokens, W, which they can allocate to a private good, X, or a public good, G, during each round of play. One token buys one unit of either good. The private good yields a return of R per unit to the individual who purchases it. The public good yields a return of V per unit to all players. The players keep the profits they have earned at the end of each round, equal for player i to:

$$\Pi_i = RX_i + VG_i + V\sum_{j\neq i} G_j \text{ with } W_i = X_i + G_i \quad (6.42)$$

The game may be played for one or more rounds. If more than one round is played, the players know at the start of the game how many rounds will be played. Also, the players allocate their tokens independently of one another during each round. They are not permitted to collude, and they learn what the other players did only after each round (or after the game concludes, in some versions of the experiment).

The returns on the goods are set so that:

$$R > V \text{ and } NV > R$$

With these returns, the pareto-optimal strategy is for all to behave cooperatively and purchase nothing but the public good each round. In this way, they maximize both the group profits and their individual profits. Cooperation is not the Nash strategy, however, given the way the experiment is set up. The Nash strategy is based on the other-things-equal assumption by each player that his or her play has no effect on the play of the other players. This is the only reasonable assumption in an independent game of this nature, and it leads to a clear-cut strategy given the returns on the private and public goods: Attempt to free ride on the goodwill of others and buy only the private good. The reason is that expected marginal profit under the Nash strategy is the partial derivative of the profit function, or

$$\partial\Pi_i/\partial X_i = R - V > 0 \quad (6.43)$$

Put differently, the only equilibrium outcome of the game in which no player would want to change his decision, other things equal, is for everyone to purchase only the private good. Furthermore, this is true whether the game is played once or repeatedly for a fixed and known number of rounds. The incentive to free ride in a one-shot game is clear. That the same incentive exists in every round of a multiround game follows from backward induction. There is a clear incentive for everyone to purchase the private good in the final round of the game. Because everyone knows this, the incentive to free ride extends to the next-to-last round and, given that, to the round before that, and so on, back to the first round. In summary, the experiments are designed to induce free riding as the rational strategy.

The results from these experiments are far different, however. The students are much more cooperative than expected. In multiround games, they typically contribute about 50% of their tokens to the public good in the first round. Cooperation does diminish as the game continues, but nowhere close to zero. After 10 rounds, students still contribute from 15% to 25% of their tokens to the public good. Furthermore, the degree of cooperation is relatively insensitive to all the following variations of the experiments:

1. The size of each group: N is usually in a range of 4−10. Mark Isaac, James Walker, and Arlington Williams ran the experiment with groups ranging from 4 to 100 and found that group size had little effect on the results. If anything, cooperation increased slightly the larger the group (Isaac et al., 1994).
2. The number of rounds: Most of the experiments are multiround games, but some students cooperate even in one-shot games.
3. Whether the subjects know the outcomes from previous rounds or not as the game progresses: The one exception was a study by Joachim Weimann, in which cooperation declined sharply when the other players were perceived to be very selfish. Weimann noted that the subjects apparently expect their cooperation to be reciprocated (Weimann, 1994).
4. Whether the same groups play each round and come to know one another or the groups are randomly reformed each round: The experiments show no evidence of reputation building; in fact, James Andreoni in his experiments found that "strangers" cooperated more than "partners" (Andreoni, 1995b).
5. The amount of the return to the public good: There does appear to be more cooperation the larger V is, but the difference is slight. And V cannot be larger than R if the incentive to free ride is to be maintained.

Kindness, Confusion, or a Warm Glow from Giving?

Andreoni conducted two separate and widely cited experiments to try to understand the motivation behind the excessive cooperation (Andreoni, 1995a,b). One experiment was designed to determine the extent to which cooperation was the result of kindness toward others or confusion about the incentive structure. He had the students play three different versions of the game, which he called the regular game, the rank game, and the regular/rank game. Each game lasted 10 rounds. The regular game is the standard game described above, in which the students keep the profits from each round of the game. The Rank game offers the students a fixed payoff that is based on the rankings of their profits over the course of the entire game. The students learn the rankings after each round. The regular/rank game is the standard regular game with one difference: The students are told their rankings after each round.

The idea behind the rank game is to place the students in a zero-sum situation that gives them absolutely no incentive to cooperate out of kindness. A student who cooperates knows that this helps the noncooperators even more. The noncooperators get their own private returns plus the public return and move ahead of the cooperators in the rankings. This becomes clear as the rankings are announced each round. Reciprocal kindness is out of the question. Therefore, Andreoni argues that any cooperation in the rank game must be the result of confusion about the nature of the game.

The only difference between the regular/rank and rank games is the method of payment. The former is a positive-sum game and the latter is a zero-sum game. Therefore, Andreoni argues that any increase in cooperation in the regular/rank game over the rank game is a measure of cooperation resulting from kindness.

Andreoni's experiments produced the expected results. The amount of cooperation in rounds 1 and 10 for each of the games was as follows:

	Percent of Tokens to the Public Good		Percent of Subjects Contributing Zero to the Public Good	
Game	Round 1	Round 10	Round 1	Round 10
Regular	56	26.5	20	45
Rank	32.7	5.4	35	92.5
Regular/ rank	45.8	9.0	10	65

The Regular game yielded the typical outcomes for these experiments. The Rank game produced a huge decrease in cooperation and students were more cooperative in the regular/rank game than in the rank game. Looking at the outcomes over all 10 rounds, Andreoni concluded that about half of the cooperation was the result of kindness and half was the result of confusion. Moreover there was a distinct change in both effects from the early rounds (1−6) to the later rounds (7−10). Throughout the early rounds, kindness increased and confusion decreased. Throughout the later rounds, kindness decreased and confusion remained fairly constant. He concluded from this that the typical pattern of decay in cooperation in the later rounds in these experiments is due to the frustration that kindness is not reciprocated. It is not the result of learning the incentive structure, which is a common explanation in the literature.

Positive versus Negative Framing

In a second experiment, Andreoni discovered that the way in which the game was framed for the students had an enormous impact on the outcome. In the standard game, the students are told by the instructions that if they invest in the public good, every member of the group benefits, and this is true no matter who invests in the public good. Andreoni refers to this instruction as positive framing because it emphasizes the benefit of doing something good. It suggests that each student is endowed with tokens of private goods and the issue for them is how many of the private tokens they will exchange for the public good to benefit everyone. Andreoni then ran a second experiment in which the students were told that if they invest in the private good, they reduce the earnings of all the other people by an amount V, the return on the public good, and this is true no matter who invests in the public good. Andreoni refers to this instruction as negative framing because it emphasizes the costs of doing something bad. The negative frame in effect rewrites the profit function, Eqn (6.42), as

$$\Pi_i = RX_i + VG_i + V\sum_{j\neq i}\left(W_j - X_j\right), \text{ or} \tag{6.44}$$

$$\Pi_i = RX_i + VG_i - V\sum_{j\neq i}X_j + VW_j\left(N-1\right) \tag{6.45}$$

It suggests that each student is endowed with his or her opponents' tokens in public goods, $V^*W_j^*(N-1)$, which endowment is lost only if they go into private goods.

The two games were identical, of course, with the same clear-cut incentive to free ride. Yet, the outcomes were quite different, with the negative frame game yielding only about half the amount of cooperation over the 10 rounds as the positive-frame game. The students apparently enjoy doing a good deed more than they enjoy not doing a bad deed.

Andreoni's previous research on charitable giving had shown that the amount and extent of charitable giving in the

United States far exceed what would be expected from altruism alone. This led him to conclude that people experience a "warm glow" from the act of giving to others in and of itself, in addition to whatever impulse they may have to be altruistic. He views these two experiments as further support for his "warm glow" hypothesis. We will return to Andreoni's research on charitable giving in Chapter 10.

Following up on Andreoni's research, Thomas Palfrey and Jeffrey Prisbrey recently made an important contribution to our understanding of the motivation behind excessive cooperation (Palfrey and Prisbrey, 1997). Their innovation was to introduce far more variation in the payoffs than in previous experiments. They ran four sessions with 10 rounds per session. The value of the public good, V, varied over the four sessions. In addition, the value of the private good, R, was determined by a random draw from a distribution in each round of every session. The variation in R was such that at times, $R < V$, giving the subjects a clear incentive to invest in the public good, and at other times, $R > NV$, giving the subjects an equally clear incentive to invest in the private good. In some sessions, the students were given one token per round; in other sessions, they were given nine tokens to test for irrational splitting of the tokens between the private and public goods each round. The variation in the payoffs allowed for a probit analysis of the results to test for heterogeneity among the subjects.

The experimental framework of Palfrey and Prisbrey allowed them to conduct the following tests:

1. *Kindness toward others (altruism):* Kindness exists if the subjects contribute more to the public good as V increases, other things being equal. They found no evidence of kindness, unlike Andreoni.
2. *A warm glow effect:* They tested for a warm-glow threshold, g, such that when $R > V$, so that the incentive was to free ride, if
   $R - V < g$, then the subjects contributed to the public good.
   $R - V > g$, then the subjects contributed to the private good.
   $R - V = g$, then the subjects contributed to either good.
   They found a warm-glow threshold in line with Andreoni's hypothesis about the motivation for excessive cooperation. But Palfrey and Prisbrey also found that the threshold varied considerably among the subjects.
3. *Gross errors as evidence of confusion:* The test for confusion was whether the subjects committed one or more of three gross errors that the structure of their experiment made possible:
   a. Splitting the tokens between the public and private good when they were given nine tokens.
   b. "Spite": contributing to the private good when $R < V$.
   c. "Sacrifice": contributing to the public good when $R > NV$.

They found evidence of these errors early on, but they virtually disappeared in the later rounds. The eventual elimination of gross errors led Palfrey and Prisbrey to conclude that the decline in cooperation over time in these free-riding experiments is most likely due to a reduction in confusion as the subjects begin to understand the game. They found no evidence of attempts to build reputation or of an increase in selfishness as the game progressed—that is, no noticeable change in the subjects' preferences.

Staged experiments must always be viewed with caution, especially when the subjects are shown to be somewhat confused by the experiments. Nonetheless, the overwhelming weight of the free-rider experiments is that people are willing to behave cooperatively even when it is clearly in their interests to behave selfishly. And, as Andreoni points out, the real world is likely to be more conducive to acts of kindness than these experimental settings are. These findings are somewhat encouraging for the mainstream normative public sector theory.

AGGREGATE EXTERNALITIES

Thus far we have considered two kinds of externalities that are likely to cause severe practical problems for the government: (1) individualized externalities arising from exclusive activities for which the identity of each individual consumer matters and (2) nonexclusive goods. Fortunately, a number of important externalities have a special form that is much more amenable to corrective public policy action.

Consider the example of highway congestion. An additional car on a congested highway generates an external diseconomy to anyone driving on the highway because it adds to the total number of vehicles on the road and to the total amount of congestion. But no one cares who is actually driving the additional car. This is an example of an *aggregate externality*, meaning that the external effect depends only upon the aggregate level of some exclusive economic activity. The identity of the individuals within the aggregate is irrelevant.

To formalize the idea of an aggregate externality, let X_{ik} = person i's driving on a particular highway, good k. Write:

$$C = C\left(\sum_{i=1}^{H} X_{ik}\right) = C(X_k), \frac{\partial C}{\partial X_{ik}} = \frac{\partial C}{\partial X_k} \quad \text{all } i = 1, \ldots, H \tag{6.46}$$

where

C = congestion on the highway.

X_k = aggregate number of cars on the highway at any given time (assuming one person per car).

The condition $\partial C/\partial X_{ik} = \partial C/\partial X_k$ implies that a decision by anyone to drive on the highway has an identical marginal effect on total congestion.

If consumers only care about the aggregate level of congestion, then they each have a utility function of the form:

$$U^h = U^{*h}[X_{hn}; X_{hk}; C(X_k)] = U^h\left(X_{hn}; X_{hk}; \sum_{i=1}^{H} X_{ik}\right) \tag{6.47}$$

where

X_{hn} = good (factor) n consumed (supplied) by person h. $n = 1, \ldots, k-1, k+1, \ldots, N$, each assumed to be a pure private good (factor).
X_{hk} = use of the highway by person h.
C and X_k, as above.

U^h ( ) has the following properties:

$$\frac{\partial U^{*h}}{\partial X_{ik}} = \frac{\partial U^h}{\partial X_k}, \quad \text{for } i \neq h \tag{6.48}$$

$$\frac{\partial U^{*h}}{\partial X_{ik}} = \frac{\partial U^h}{\partial X_{hk}} + \frac{\partial U^h}{\partial X_k}, \quad \text{for } i = h \tag{6.49}$$

If anyone other than person h uses the road, his or her utility is affected simply because aggregate road use increases, thus increasing congestion. When person h uses the road, however, there are two distinct effects. On the one hand, person h has some private reason for choosing to drive on the road that is unrelated to the congestion problem. On the other hand, he or she is adding to the congestion exactly as any other driver would and with the same consequences for his or her utility. He or she may or may not consciously understand that his or her choice to drive on the road necessarily contributes to the congestion and thereby lowers his or her utility (a point we will return to later), but he or she certainly views his or her own use of the road differently from anyone else's use of the road. This is why the derivative of U^h with respect to X_{hk} has two separate terms: a private-use term and a congestion term. Note, finally, that congestion must be a function of aggregate road use and not a general function of individual road use such as $C^* = C^*(X_{1k}, \ldots, X_{ik}, \ldots, X_{Hk})$. With this more general formulation, $\partial U^{*h}/\partial U_{ik} \neq \partial U^{*h}/\partial X_{jk}$ for $i \neq j$, and we are back in a situation of individualized externalities in which the identity of the individual consumer matters.

Congestion is not the only important example of an aggregate externality by any means. Virtually all pollution externalities affecting consumers, whether caused by other consumers or by producers, can be thought of as aggregate externalities arising from exclusive economic activities. Smog, airport noise, and industrial air and water pollution usually exhibit this property.

The Pigovian Tax

Aggregate externalities are far more amenable to government policy than are the individualized externalities or nonexclusive goods. The government need not design a set of H taxes, one specific to each individual. They can be corrected by a single tax levied on the externality-causing activity. The single tax solution requires one additional behavioral assumption, that when an individual engages in the activity for his or her own personal reasons, he or she ignores the effect of his or her activity on the aggregate externality. This is certainly a plausible assumption.

To derive the single tax result, consider social welfare maximization when a single exclusive good (factor) X_k gives rise to an aggregate externality affecting all consumers. Assume all other $(N-1)$ goods and factors are purely private. The government's problem becomes

$$\max_{(X_{hn};\, X_{ik})} W\left[U^h\left(X_{hn}; X_{hk}; \sum_{i=1}^{H} X_{ik}\right)\right]$$
$$\text{s.t. } F\left(\sum_{h=1}^{H} X_{hn}; \sum_{i=1}^{H} X_{ik}\right) = 0$$

where

$n = 1, \ldots, k-1, k+1, \ldots, N$.

The corresponding Lagrangian equation is

$$\max_{(X_{hn};X_{ik})} L = W\left[U^h\left(X_{hn}; X_{hk}; \sum_{i=1}^{H} X_{ik}\right)\right] + \lambda F\left(\sum_{h=1}^{H} X_{hn}; \sum_{i=1}^{H} X_{ik}\right)$$

Interpersonal Equity Conditions

As always in the first-best analysis of consumer externalities, we need the interpersonal equity to obtain the pareto-optimal conditions. Consider the first-order conditions with respect to two different people's consumption (supply) of good 1, say X_{hl} and X_{jl}. The first-order conditions are

$$X_{h1}: \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{h1}} = -\lambda F_1 \tag{6.50}$$

$$X_{j1}: \frac{\partial W}{\partial U^j}\frac{\partial U^j}{\partial X_{j1}} = -\lambda F_1 \tag{6.51}$$

Thus,

$$\frac{\partial W}{\partial U^i}\frac{\partial U^i}{\partial X_{i1}} = -\lambda F_1 \quad \text{all } i = 1,\ldots,H \tag{6.52}$$

the standard result.

Pareto-Optimal Conditions

To derive the efficiency conditions, compare the first-order conditions for person i's consumption of the externality good (factor) k and his or her consumption of any other good (factor), say good 1 (X_{i1}). The first-order conditions are

$$X_{ik}: \frac{\partial W}{\partial U^i}\frac{\partial U^i}{\partial X_{ik}} + \sum_{h=1}^{H}\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{ik}} = -\lambda F_k \tag{6.53}$$

$$X_{i1}: \frac{\partial W}{\partial U^i}\frac{\partial U^i}{\partial X_{i1}} = -\lambda F_1 \tag{6.54}$$

Condition (6.53) for X_{ik} reflects both the personal enjoyment that person i receives from good (factor) k (the first term) and the externality from his or her consumption that affects everyone, *including himself* (second term). Because the externality is of the aggregate form, condition (6.53) can be rewritten:

$$\frac{\partial W}{\partial U^i}\frac{\partial U^i}{\partial X_{ik}} + \sum_{h=1}^{H}\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_k} = -\lambda F_k \tag{6.55}$$

when

$$X_k = \sum_{i=1}^{H} X_{ik}$$

Next, follow the usual procedure for obtaining the pareto-optimal conditions by dividing Eqn (6.55) by Eqn (6.54) to obtain

$$\frac{\frac{\partial W}{\partial U^i}\frac{\partial U^i}{\partial X_{ik}} + \sum_{h=1}^{H}\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_k}}{\frac{\partial W}{\partial U^i}\frac{\partial U^i}{\partial X_{i1}}} = \frac{F_k}{F_1} \quad i = 1,\ldots,H \tag{6.56}$$

Assuming the interpersonal equity conditions have been achieved for good (factor) 1, separate the LHS of Eqn (6.54) into $(H+1)$ terms, selectively substitute the interpersonal equity conditions term by term to match the marginal social welfare terms, $\partial W/\partial U^h$ in the numerator and denominator, and cancel the social welfare terms to yield

$$\frac{\frac{\partial U^i}{\partial X_{ik}}}{\frac{\partial U^i}{\partial X_{i1}}} + \sum_{h=1}^{H}\left(\frac{\frac{\partial U^h}{\partial X_k}}{\frac{\partial U^h}{\partial X_{h1}}}\right) = \frac{F_k}{F_1} \quad i = 1,\ldots,H \tag{6.57}$$

Condition (6.55) can be written as

$$\text{MRS}^i_{X_{ik},X_{i1}} + \sum_{h=1}^{H}\text{MRS}^h_{X_k,X_{h1}} = \text{MRT}_{X_k,X_1} \quad i = 1,\ldots,H \tag{6.58}$$

Pareto optimality requires that the MRT between goods (factors) k and 1 be equal to each person's MRS between his or her personal use of k and good 1, plus the summation of everyone's (his or her own included) MRS between the externality and good 1. Bringing all the externality terms over to the RHS,

$$\text{MRS}^i_{X_{ik},X_{i1}} = \text{MRT}_{X_k,X_1} - \sum_{h=1}^{H}\text{MRS}^h_{X_k,X_{h1}} \quad i = 1,\ldots,H \tag{6.59}$$

Notice that the RHS of Eqn (6.59) is independent of i. That is, each consumer's "personal-use" MRS differs from the MRT by the same amount, the summation of all the marginal external effects. This differs significantly from the result when the externality depends upon who consumed good k, the individualized externality. In that case, the required pareto optimality is

$$\sum_{h=1}^{H}\text{MRS}^h_{X_{ik},X_1} = \text{MRT}_{X_k,X_1} \tag{6.60}$$

or

$$\text{MRS}^i_{X_{ik},X_{i1}} = \text{MRT}_{X_k,X_1} - \sum_{h\neq i}\text{MRS}^h_{X_{ik},X_{h1}} \tag{6.61}$$

Hence, the personal-use marginal rates of substitution differ from the MRT by a variable amount, depending upon whose personal use is being considered. Consequently, H Pigovian taxes are required to correctly allocate good k, one for each consumer.

In contrast, only a single Pigovian tax is necessary in the aggregate case. Let good (factor) 1 be the numeraire, $P_1 \equiv 1$. Faced with a producer price P_k, the producers set $P_k = \text{MRT}_{X_k,X1}$ by profit maximization. Faced with a consumer price q_k, each person consumes good k such that his or her personal use $q_k = \text{MRS}^i_{X^i_{ik},X_{i1}}$ $i = 1,\ldots,H$, assuming that he or she ignores the marginal external effect of his or her consumption (supply). Therefore, to achieve pareto optimality, place a unit tax, t_k, on each consumer equal to $-\sum_{h=1}^{H}\text{MRS}^h_{X_k,X_{h1}}$, the sum of the marginal external effects. With the unit tax and assuming utility and profit maximization,

$$q_k = P_k + t_k \tag{6.62}$$

and

$$\text{MRS}^i_{X_{ik},X_{i1}} = \text{MRT}_{X_k,X_1} - \sum_{h=1}^{H}\text{MRS}^h_{X_k,X_{h1}} \tag{6.63}$$

as required. If the external effects are diseconomies, such as congestion, $t_k > 0$, following the convention that MRS > 0 for goods, MRS < 0 for bads. External economies are subsidized ($t_k < 0$).

Note that the single Pigovian tax is correct only under two conditions: (1) the externality has a simple, aggregate formulation and (2) consumers ignore all external effects when maximizing utility. The behavioral assumption is crucial because if any consumer considers so much as a single external effect, the single Pigovian tax is no longer pareto optimal. Suppose, for example, that consumer i considered both the direct personal effect and the indirect externality effect on himself or herself when deciding how much of good k to consume. He or she would then equate the gross of tax price, q_k, to $\mathrm{MRS}^i_{X_{ik},X_{i1}} + \mathrm{MRS}^i_{X_k,X_{i1}}$ to maximize utility, and the single tax scheme breaks down. The government would need an additional tax for each consumer who considered external effects in this way, and the aggregate externality would be just as difficult to correct as an individualized externality.

The aggregate externality case is easy to represent with standard supply and demand analysis. In Fig. 6.6, S_k is a normal supply curve, representing the marginal costs of X_k (the MRT in terms of good 1, the numeraire). D^p_k is the "private" aggregate demand curve, obtained by horizontal summation of the individuals' personal-use demand curves reflecting their personal-use marginal rates of substitution. D^{soc}_k is the true "social" demand curve equal, at every X_k, to D^p_k plus the (negative) aggregate marginal external effects, $\sum_{h=1}^{H} \mathrm{MRS}^h_{X_k,X_{h1}}$. The Pigovian tax forces consumers onto D^{soc}_k, establishing the pareto-optimal equilibrium at the intersection of D^{soc}_k and S_k.

Finding the Optimum by Trial and Error

Figure 6.6 highlights an important property of the tax: It must equal the sum of the marginal external effects

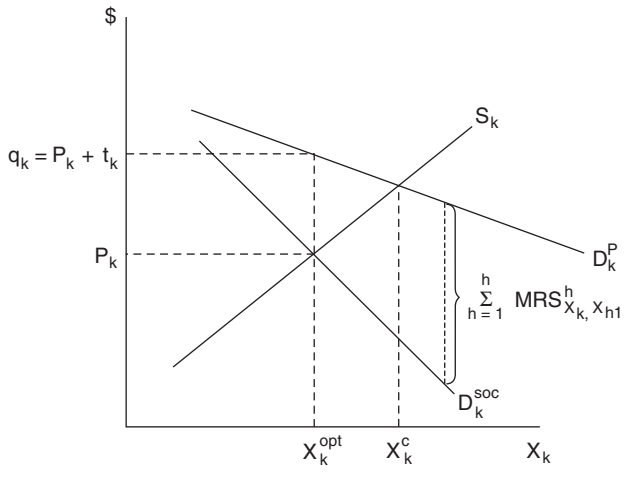

FIGURE 6.6

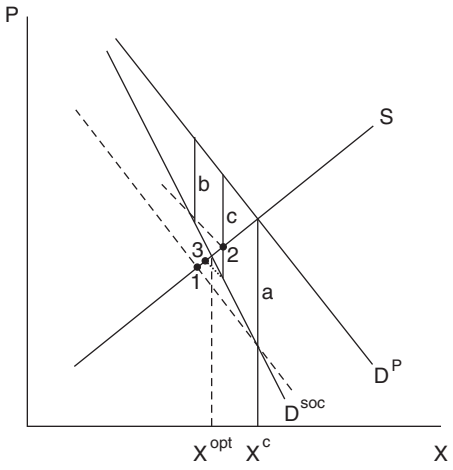

FIGURE 6.7

$(-\sum_{h=1}^{H} \mathrm{MRS}^h_{X_k,X_{h1}})$ at the optimal level of X_k. Setting the tax equal to the aggregate marginal damage at the initial output X^c_k, the competitive equilibrium without the tax, is not correct.

Given that only a single tax is necessary, however, the government may be able to reach the correct tax (approximately) by trial and error even if its initial choice is incorrect. The effectiveness of any trial and error solution depends upon four factors: the nature of the trial and error process used, the government's ability to assess aggregate marginal damages, the shape of the marginal damage function ($\sum_{h=1}^{H} \mathrm{MRS}^h_{X_k,X_{h1}}$, the difference between D^p and D^{soc} at each X_k), and the stability of the competitive market being taxed.

Refer to Fig. 6.7. Assume the curves S, D^p, and D^{soc} in the figure accurately describe the competitive market for some activity and the aggregate marginal damages stemming from the activity. The following trial and error process is stable and generates t^{opt}_k in the limit.[13] The government sets an initial tax equal to the marginal damages at the no-tax equilibrium and recomputes the tax to equal the marginal damages at each successive equilibrium. The resulting pattern of equilibria converge to X^{opt}_k.

The tax, t_k, is initially set at a, equal to the marginal damages at X^c, the no-tax competitive equilibrium. With $t_k = a > t^{opt}_k$, the market overshoots X^{opt}, establishing a new equilibrium at point 1 on S. The marginal damages have been reduced to b, however, so t_k is adjusted to equal b. This tax overshoots X^{opt} in the opposite direction, bringing the economy to point 2 on S. Readjusting the tax

13. See Kraus and Mohring (1975), Baumol (1972), for further discussion of the suitability of sequential pollution taxes in determining a global optimum.

to equal c, the new higher marginal damages, brings the economy to point 3 on S, and so forth. The trial and error process approaches X^{opt} in the limit.

The trial and error process works in this market because it is stable and the marginal damages are positively related to the level of economic activity. Most markets with externalities are likely to have the same properties. Therefore, it is reasonable to assume that simple trial and error processes can generate results that are at least approximately optimal for a broad range of aggregate externalities.[14]

Two Caveats to the Pigovian Tax

The Pigovian single tax solution comes with two caveats. The first caveat is the usual one of all first-best analysis. If the government cannot achieve the interpersonal equity conditions by means of lump-sum redistributions of income, and there is no reason to suppose that it can, then a tax equal to $-\sum_{h=1}^{H} \text{MRS}^{h}_{X_k, X_{h1}}$ may not be consistent with the (constrained) social optimum. We will return to this point in Chapter 20, which discusses externality theory in a second-best framework.

The second caveat is a more narrow distributional point. Optimally correcting for an aggregate externality with a Pigovian tax is potentially pareto superior to the initial situation without the tax. Everyone can be made better off by moving to the first-best utility–possibilities frontier from an inefficient point below the frontier. But whether everyone actually is better off with the Pigovian tax depends on what the government does with the tax revenues collected. The highway congestion example is a good case in point. The Pigovian tax is supposed to benefit the drivers on the congested highway, but the drivers could be made worse off if the revenues are not returned to them, in which case the very people the government is trying to help with the tax will oppose it.

Figure 6.8 illustrates, as in Fig. 6.6, D^p is the private market demand curve, reflecting only the private-use value of driving on the highway. D^{soc} is the social demand curve; it lies below D^p at every output by the aggregate losses to the drivers on the margin resulting from the congestion. The supply curve, S_k, assumes constant marginal cost of P_k to focus on the drivers' problem. The optimal Pigovian tax is t_k. Without a tax, the competitive equilibrium is (X_k^c, P_k), at the intersection of S_k and D^p. With the optimal Pigovian tax, the equilibrium road use drops to X_k^{opt}, at the intersection of S_k and D^{soc}. The price to the drivers rises to $P_k + t_k$, and the tax revenue collected from them is $t_k X_k^{opt}$.

Assume no income effects so that consumer surplus is an appropriate income measure of the drivers' welfare. The

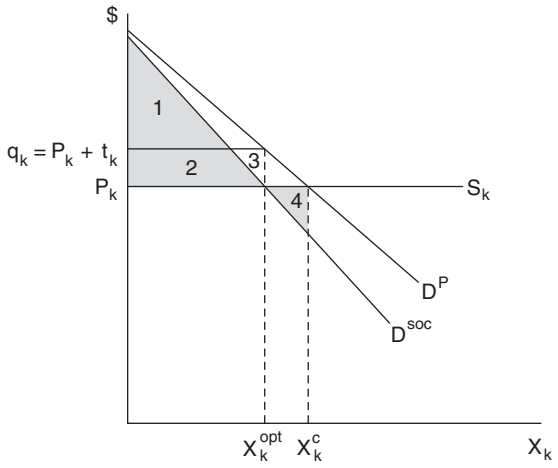

FIGURE 6.8

potential consumer surplus at any output is the area between D^{soc} and S_k to that output. At the no-tax equilibrium X_k^c, the drivers' consumer surplus equals areas $1 + 2 - 4$. Area 4 represents the loss caused by excessive congestion at the no-tax equilibrium. At the pareto-optimal output X^{opt}, the potential consumer surplus available to the drivers equals area $1 + 2$, but the drivers obtain this surplus only if the tax revenue, equal to area $2 + 3$ is returned to them. The drivers are clearly better off at X_k^{opt} if they receive the tax revenue; they avoid the excessive congestion at the no-tax equilibrium, represented by area 4. If the tax revenue is not returned, however, the drivers' actual consumer surplus is only area $1-3$. Whether they are now better off at X^{opt} depends on the relative size of areas $1 + 2 - 4$ and $1-3$. If the tax revenue $(2 + 3)$ exceeds area 4, the drivers are worse off at the optimum and they will resist the tax.[15]

This analysis may explain why commuters tend to resist tolls that are intended to reduce highway congestion by diverting some of them to other means of transportation. The commuters know that they will not receive the toll revenue. In their view, they will simply face higher commuting costs that exceed the value to them of the reduced congestion.

We should note that this second caveat is not entirely consistent with the first-best policy assumptions. First-best analysis assumes that the government engages in allocational policies to bring society to the first-best utility possibilities frontier *and* that it redistributes lump-sum to reach the bliss point on the frontier. The caveat ignores the distributional part of the policy. Whether the drivers are better or worse off at the bliss point ultimately depends on society's social welfare rankings and the interpersonal equity conditions that are derived from them. The disposition of Pigovian tax revenues may be taken into

14. There are other means besides taxes for achieving the optimum. We will consider some of the alternatives in Chapter 7 in the context of a production externality.

15. We were made aware of this caveat by Russell Roberts in a seminar that he gave at Boston College.

consideration by the government when it redistributes, but it is irrelevant to determining the final distribution of income. Nonetheless, resistance to tolls and other forms of externality taxes is quite vocal, perhaps because people do not believe that the government has a fully articulated distributional policy. Therefore, they react more to their direct gains and losses from the government's allocational policies than to the efficiency gains from the policies.

REFERENCES

Andreoni, J., February 1995a. Warm glow vs. Cold prickly: the effects of positive and negative framing on cooperation experiments. Quarterly Journal of Economics 110 (1), 1–22.

Andreoni, J., September 1995b. Cooperation in public-goods experiments: kindness or confusion? American Economic Review 85 (4), 891–904.

Arrow, K.J., 1977. The organization of economic activity: issues pertinent to the choice of market versus nonmarket allocation. In: Haveman, R., Margolis, J. (Eds.), Public Expenditure and Policy Analysis, second ed. Rand McNally College Publishing, Chicago.

Baumol, W.J., June 1972. On taxation and the control of externalities. American Economics Review 62 (3), 307–322.

Clarke, E.H., Fall 1971. Multipart pricing of public goods. Public Choice 11, 17–33.

Clarke, E.H., 1972. Multipart pricing of public goods: an example. In: Mushkin, S. (Ed.), Public Prices for Public Products. Urban Institute, Washington, D. C.

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

Groves, T., Loeb, M., August 1975. Incentives and public inputs. Journal of Public Economics 4 (3), 211–226.

Isaac, M., Walker, J., Williams, A., May 1994. Group size and the voluntary provision of public goods. Journal of Public Economics 54 (1), 1–36.

Johansen, L., September 1963. Some notes on the lindahl theory of determination of public expenditures. International Economic Review 4 (3), 346–358.

Kraus, M., Mohring, H., May 1975. The role of pollutee taxes in externality problems. Economica 42 (166), 171–176.

Lindahl, E., 1958. Die Gerechtigkeit der Besteverung. Lund, 1919, reprinted (in part). In: Musgrave, R., Peacock, A. (Eds.), Classics in the Theory of Public Finance. Macmillan, London.

Mill, J.S., 1921. In: Ashley, W.J. (Ed.), Principles of Political Economy. Longmans Green & Co., London.

Ng, Y.-K., April 1975. The paradox of universal externality. Journal of Economic Theory 10 (2), 258–264.

Pigou, A.C., 1932. The Economics of Welfare, fourth ed. Macmillan, London.

Palfrey, J., Prisbrey, J., December 1997. Anomalous behavior in public goods experiments: how much and why? American Economic Review 87 (5), 829–846.

Samuelson, P.A., November 1954. The pure theory of public expenditure. Review of Economics and Statistics 36 (4), 387–389.

Samuelson, P.A., November 1955. Diagrammatic exposition of a theory of public expenditure. Review of Economics and Statistics 37 (4), 350–356.

Samuelson, P.A., November 1958. Aspects of public expenditure theories. Review of Economics and Statistics 40 (4), 332–338.

Samuelson, P.A., 1969. Pure theory of public expenditures and taxation. In: Guitton, H., Margolis, J. (Eds.), Public Economics. St. Martin's Press, New York.

Smith, A., 1904. In: Cannan, E. (Ed.), The Wealth of Nations, vol. II. G. P. Putnam's Sons, New York, p. 310.

Tideman, T.N., Tullock, G., December 1976. A new and superior process of making social choices. Journal of Political Economy 84 (6), 1145–1159.

Weimann, J., June 1994. Individual behavior in a free riding experiment. Journal of Public Economics 54 (2), 185–200.

Chapter 7

Production Externalities

Chapter Outline

A policy-relevant, technological production externality has two properties: Production activity by some firm directly enters into (or "alters") the production function of at least one other firm, and the external effect is not captured in the marketplace. These properties are completely analogous to those of a policy-relevant, technological consumption externality. Therefore, having analyzed various consumption externality models in some detail, the treatment of production externalities can be fairly brief. The production models and the resulting pareto-optimal decision rules for production externalities are virtually identical to their consumption counterparts, with the roles of consumption and production reversed. In particular, there are these important similarities:

1. The pareto-optimal decision rules for consumption externalities require equating marginal rates of transformation in production to summations of marginal rates of substitution in consumption. For production externalities, summations of marginal rates of transformation in production (alternatively, marginal rates of technical substitution or marginal products) equal marginal rates of substitution in consumption.
2. In both instances, the government can achieve pareto optimality by retaining decentralized markets and taxing (subsidizing) an externality-generating exclusive activity.
3. We saw that public policy is problematic in the case of individualized consumption externalities because the government must design a set of H corrective taxes, one for each of H people consuming the good. In contrast, when the external effect depends only on aggregate consumption, a single tax paid by all consumers can achieve the pareto-optimal conditions. The same differences in tax policies apply to production externalities.

Because of these similarities, this chapter presents only the aggregate production externality model. The aggregate model is by far the one most widely used in policy applications, and it provides a simple analytical framework for considering a number of policy implications that could have been discussed in the preceding chapter but are especially intuitive in a production framework. Most of the policy examples in this chapter center on pollution control, as industrial pollution is a particularly appropriate and important application of the aggregate production externalities model. Chapter 8 then discusses global warming as an extended example incorporating both production and consumption externalities.

Having analyzed aggregate production externalities and noted their similarities with aggregate consumption externalities, the reader should have no difficulty modeling other types of production externalities. The other production

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00007-8

cases are also closely analogous to their consumption counterparts.

THE CONDENSED MODEL FOR PRODUCTION EXTERNALITIES

The analysis of consumption externalities used a condensed version of the general equilibrium model in Chapter 2 for its analytical framework of the form

$$\max_{(X_{ik})} W\left[U^h()\right]$$
$$\text{s.t.} F\left(\sum_{i=1}^{H} X_{ik}\right) = 0$$

where X_{ik} was defined as the consumption of good k by person i. The way in which the X_{ik} entered each person's utility function determined the appropriate policy response by the government.

Production externalities can also be analyzed with a condensed version of the full general equilibrium model, the only difference being that the model must highlight possible interdependencies in production rather than in consumption. To achieve this, we will ignore once again any notational distinction between goods and factors but define the arguments, X, in terms of production. Let X_{ji} = good (factor) i supplied (demanded) by firm j, with factors measured negatively, $j = 1,\ldots, J$ and $i = 1,\ldots, N$. There are J firms and N goods and factors.

Since we are now interested in production interrelationships, writing production as a single production-possibilities frontier is no longer useful. The model must retain the individual-firm production functions. Define $f^k() = 0$ as the implicit production function for firm k, $k = 1,\ldots, J$. Write

$$f^k\left(X_{ji}\right) = 0 \quad k = 1,\ldots,J \tag{7.1}$$

as the most general notation. This allows for the worst possible case of individualized externalities, in which each of the J production relationships has JN arguments: The production (use) of any of the N goods (factors) by any of the J firms in the economy affects every firm. In this model, each firm could produce multiple outputs, rather than a single output as in the Chapter 2 model. The model also permits each good and factor to be produced, although this is not necessary. J can be larger or smaller than N.[1]

Analogous with consumption externalities, define a pure public good (factor) as one for which

$$\frac{\partial f^k}{\partial X_{ji}} \equiv f^k_{ji} \neq 0 \quad \text{all } k,j = 1,\ldots,J \tag{7.2}$$

That is, production (use) of good (factor) i affects all production relationships on the margin no matter where activity i occurs. This is the worst case described above. Similarly, a pure private good (factor) is one for which

$$\frac{\partial f^k}{\partial X_{ji}} \equiv f^k_{ji} = 0 \quad k \neq j \tag{7.3}$$

Firm k's use or production of i affects only itself on the margin. Production with private goods and factors is represented notationally as $f^k(X_{ki}) = 0$, analogous with the notation of Chapter 6.

The condensation occurs in the household sector of the Chapter 2 model. Interrelationships among consumers are irrelevant to the study of production externalities, so that it is no longer necessary to retain a many-consumer economy along with the social welfare function to resolve distributional questions. These could be retained, to be sure, but the existence of production externalities does not alter any of the pareto-optimal consumption conditions or the interpersonal equity social welfare conditions that are necessary for reaching the first-best bliss point. No loss of generality occurs, then by assuming a one-consumer equivalent economy in which the consumer supplies all factors of production and receives all the produced goods and services, providing it is understood that one-consumer equivalence arises because the government is optimally redistributing lump sum to satisfy the interpersonal equity conditions of social welfare maximization. Without this assumption (or one of the severe restrictions on preferences that are sufficient for one-consumer equivalence), the pareto-optimal conditions developed in this chapter would literally apply only to an economy with one consumer. They would not have any normative policy significance.

With this understanding, the household sector of the model can be represented as

$$U(X_1;\ldots,X_i,\ldots,X_N) = U(X_i) \tag{7.4}$$

where X_i = aggregate production of (demand for) good (factor) i. Finally, market clearance implies

$$X_i = \sum_{j=1}^{J} X_{ji} \quad i = 1,\ldots,N \tag{7.5}$$

Equations (7.5) can be incorporated directly into the utility function as

$$U = U\left(\sum_{j=1}^{J} X_{ji}\right) \tag{7.6}$$

with the understanding that

$$\frac{\partial U}{\partial X_{ji}} \frac{\partial U}{\partial X_i} = U_i \quad j = 1,\ldots,J; \text{ all } i = 1,\ldots,N$$

1. J is much larger than N in actual economies—the number of firms far exceeds the number of goods and factors.

That is, the consumer does not care where the production activity occurs.

Thus, the complete general model for analyzing production externalities is

$$\max_{(X_{ik})} U\left(\sum_{i=1}^{H} X_{ik}\right)$$
$$\text{s.t.} f^k(\,) = 0 \quad k = 1, \ldots, J$$

The arguments of the individual production functions $f^k(\,)$ depend on the exact form of the production externality.

AGGREGATE PRODUCTION EXTERNALITIES

Industrial water pollution offers an appropriate context for the analysis of the aggregate externality case. Suppose that all firms are located on the shore of a lake and that they all use the water as a coolant for their production processes. Using the water in this manner heats it up, so that each firm returns the water to the lake at a higher temperature than it was originally received. The hotter the water, the less effective it is as a cooling agent. The heat, then, is the source of a technological production externality (a diseconomy), because each firm's production function is directly affected. Furthermore, suppose the firms do not care who is heating the water. All that matters is the amount that the water temperature increases, which is a function only of the total amount of water used by all the firms as a cooling agent. The heat pollution is an example of an aggregate externality.[2]

To model this example, let factor i be water and assume that all other goods and factors are purely private. The production relationships in this case are

$$f^{*k}(X_{kn}; X_{ki}; H) = 0 \quad n = 1, \ldots, i-1, i+1, \ldots, N \quad (7.7)$$
$$K = 1, \ldots, J$$

with

$$H = H\left(\sum_{j=1}^{J} X_{ji}\right) \quad (7.8)$$

where $H =$ the water temperature, and $\frac{\partial H}{\partial X_{ji}} = \frac{\partial H}{\partial X_i}$, all $j = 1, \ldots, J$

Substituting for H in f^{*k} yields

$$f^k\left(X_{kn}; X_{ki}; \sum_{j=1}^{J} X_{ji}\right) = 0 \quad k = 1, \ldots, J \quad (7.9)$$

These production relationships distinguish between each firm's private use of water as a coolant, represented by the argument X_{ki}, and the external effect of the heat, represented by the argument $\sum_{j=1}^{J} X_{ji}$. Thus,

$$\frac{\partial f^{*k}}{\partial X_{ji}} = \frac{\partial f^k}{\partial X_{ji}} \quad j \neq k \quad (7.10)$$

$$\frac{\partial f^{*k}}{\partial X_{ji}} = \frac{\partial f^k}{\partial X_{ki}} + \frac{\partial f^k}{\partial X_i} \quad j = k \quad (7.11)$$

When some other firm uses water, firm k is affected on the margin only because the water temperature has increased. When firm k uses water, its production function is twice affected on the margin, once by the cooling effect of the water and once by the increased heat to which it contributes.

Combining Eqns (7.6) and (7.9), the complete model of social welfare maximization is

$$\max_{(X_{jn}; X_{ji})} U\left(\sum_{j=1}^{J} X_{jn}, \sum_{j=1}^{J} X_{ji}\right)$$
$$\text{s.t.} f^k\left(X_{kn}; X_{ki}; \sum_{j=1}^{J} X_{ji}\right) = 0$$
$$n = 1, \ldots, i-1, \ldots, i+1, \ldots, N \quad k, j = 1, \ldots, J$$

Supplying Lagrangian multipliers λ^k for each of the production functions, the Lagrangian is

$$\max_{(X_{jn}; X_{ji})} L = U\left(\sum_{j=1}^{J} X_{jn}, \sum_{j=1}^{J} X_{ji}\right) + \sum_{k=1}^{J} \lambda^k f^k\left(X_{kn}; X_{ki}; \sum_{j=1}^{J} X_{ji}\right)$$

The First-Order Conditions—Pareto Optimality

Production models of this type, with one-consumer equivalent economies, generate only the pareto-optimal conditions necessary to bring the economy to its first-best production possibilities frontier. They are derived by considering any two activities by any one firm. Let us first establish the important result that the presence of production externalities in some markets implies intervention only in those markets. The perfectly competitive allocation is correct for all other activities. To see this, consider the purely private goods (factors) m and 1 supplied

2. Notice that if the firms were situated along a river, as is often the case, the aggregate model would not apply. The firm farthest upstream would be unaffected by how any of the remaining firms use the water; the second firm would be affected only by the first firm's use of the water; and so on, so that it matters to each firm who uses the water. Unfortunately, industrial water and air pollution sometimes do take the form of individualized externalities, in which case the optimal public policy becomes much more difficult to implement, as we have seen with consumption externalities.

(demanded) by firm j, X_{jm}, and X_{j1} for $m \neq i$. The first-order conditions are

$$X_{jm}\colon \frac{\partial U}{\partial X_m} = -\lambda^j \frac{\partial f^j}{\partial X_{jm}} \quad \text{all } j = 1, \ldots, J \text{ any } m \neq i \tag{7.12}$$

$$X_{j1}\colon \frac{\partial U}{\partial X_1} = -\lambda^j \frac{\partial f^j}{\partial X_{j1}} \quad \text{all } j = 1, \ldots, J \tag{7.13}$$

Dividing Eqn (7.12) by Eqn (7.13)

$$\frac{\frac{\partial U}{\partial X_m}}{\frac{\partial U}{\partial X_1}} = \frac{\frac{\partial f^j}{\partial X_{jm}}}{\frac{\partial f^j}{\partial X_{j1}}} \equiv \frac{f^j_{jm}}{f^j_{j1}} \quad \text{all } j = 1, \ldots, J \tag{7.14}$$

This is the standard competitive result. The left-hand side (LHS) is the MRS between m and 1. There are three possible interpretations of the production derivatives, depending on whether m and 1 are goods or factors. Totally differentiating $f^j(\,) = 0$ with respect to X_{j1} and X_{jm} yields

$$\frac{f^j_{jm}}{f^j_{j1}} = -\frac{\mathrm{d}X_{j1}}{\mathrm{d}X_{jm}} \tag{7.15}$$

with all other goods and factors constant.

If both m and 1 are goods, the ratio is their marginal rate of transformation. If both are factors, the ratio is their marginal rate of technical substitution in production. Finally, if 1 is a good and m a factor, the ratio is the marginal product of factor m in producing good 1 (recall that factors are measured negatively). Since 1 and m can be goods or factors, conditions (7.14) reproduce pareto-optimal conditions P4 to P8 from the full model of Chapter 2. We will refer to the ratio generally as a marginal rate of transformation throughout Chapter 7 and switch to one of the other interpretations when a specific example warrants it.

To derive the pareto-optimal rules for factor i (water), which generates the aggregate externality, consider the use of water by firm j and its supply of good 1, X_{j1} and X_{j1} (assume X_1 is a good for purposes of interpretation). The first-order conditions are

$$X_{ji}\colon \frac{\partial U}{\partial X_i} = -\lambda^j \frac{\partial f^j}{\partial X_{j1}} - \sum_{k=1}^{J} \lambda^k \frac{\partial f^k}{\partial X_i} = -\lambda^j f^j_{ji} - \sum_{k=1}^{J} \lambda^k f^k_i \tag{7.16}$$

$$X_{j1}\colon \frac{\partial U}{\partial X_1} = -\lambda^j \frac{\partial f^j}{\partial X_{j1}} = -\lambda^j f^j_{ji} \quad j = 1, \ldots, J \tag{7.17}$$

Dividing Eqn (7.16) by Eqn (7.17)

$$\frac{\frac{\partial U}{\partial X_i}}{\frac{\partial U}{\partial X_1}} = \frac{\lambda^j f^j_{ji} + \sum_{k=1}^{J} \lambda^k f^k_i}{\lambda^j f^j_{j1}} \tag{7.18}$$

The LHS has a standard interpretation as the marginal rate of substitution between the consumption of good 1 and the supply of factor i (water). To interpret the right-hand side (RHS), the λ^k multipliers must be removed. To do this, note that

$$\frac{\partial U}{\partial X_{j1}} = \frac{\partial U}{\partial X_1} = -\lambda^j f^j_{ji} \quad \text{all } j = 1, \ldots, J \tag{7.19}$$

from the first-order conditions. Equation (7.19) says that the marginal "kick" to utility from the production of good 1 must be the same no matter which firm produces it. This condition holds automatically under the assumption that the consumer is indifferent to the identity of the firms. Using this result, the RHS can be cleared of the λ^k terms by separating the RHS into $J+1$ terms, making the appropriate substitution for $\lambda^j f^j_{j1}$ in the denominators to match up the corresponding λ^k in the numerators, and canceling each λ term by term. This procedure is analogous to the one used to simplify expressions for consumer externalities, with one important difference. For the consumer case, the procedure was legitimate only under the assumption that the proper lump-sum redistributions were carried out to satisfy the interpersonal equity conditions of social welfare maximization. In the production case, all that matters is that the consumer does not care which firm supplies (uses) a good (factor).[3]

Having applied this procedure, the first-order conditions become

$$\frac{\frac{\partial U}{\partial X_i}}{\frac{\partial U}{\partial X_1}} = \frac{f^j_{ji}}{f^j_{j1}} + \sum_{k=1}^{j} \left(\frac{f^j_{jm}}{f^j_{j1}} \right) \quad \text{all } j = 1, \ldots, J \tag{7.20}$$

The marginal rate of substitution between good 1 and factor i in consumption must equal, for each firm, the private-use marginal product of factor i in the production of good 1 (the cooling property of the water) plus the additional aggregate marginal effect that increased use of factor i has on the production of good 1 through the externality (i.e., the combined adverse effects on every firm's production of good 1 resulting from the increased water temperature). For firm k, the ratio $f^k_i / f^k_{k1} = -\mathrm{d}X_{k1}/\mathrm{d}H$, the (negative) marginal product of heat on its production of good 1. These two effects combined are the true social marginal product of factor i in the production of good 1. For purposes of further discussion, rewrite the condition (7.20) as

$$\mathrm{MRS}_i, 1 = \mathrm{MP}^j_{ji,j1} + \sum_{k=1}^{J} \mathrm{MP}^k_{i,k1} \quad j = 1, \ldots, J \tag{7.21}$$

3. Recall, however, that we are implicitly assuming that the interpersonal equity conditions are satisfied in specifying a one-consumer equivalent economy. In addition, all production functions are assumed to be continuous, twice differentiable, and well behaved in that their Hessians are negative definite, with all goods and factors infinitely divisible. Notice that our specification of production assumes away intermediate products.

The Pigovian Tax

Consistent with our analysis of an aggregate consumption externality, suppose that each firm considers only the private cooling properties of water when deciding how much to use. It ignores the external heat affect, not only on all others but also on itself. Under this assumption, the government can achieve pareto-optimal condition (7.21) by retaining a decentralized market for factor i and setting a unit tax on the use of i equal to the sum of its external effects on the margin. Define consumer prices q_i and q_1, producer prices p_i and q_1, and a tax t_i such that

$$\frac{q_i}{q_1} = \frac{p_i}{q_1} + \frac{t_i}{q_1} \tag{7.22}$$

(We assumed that good 1 was the numeraire when analyzing consumption externalities. Here we choose to retain the price q_1 because it often aids in the interpretation of production externalities.) The consumer sets $q_i/q_1 = \text{MRS}_{i,1}$. Each firm sets $p_i/q_1 = \text{MP}^j_{ji,j1}$, its private-use marginal product. Alternatively, $p_i = \text{MP}^j_{ji,j1} \cdot q_1$, which says that firms equate the price of an input to the value of its marginal product. This assumes, of course, that each firm ignores the external effects of using factor i. Without any government intervention, $p_i = q_i$, and the $\text{MRS}_{i,1}$ would equal the private marginal product for each firm in equilibrium. To achieve the correct pareto-optimal conditions, the government must set $t_i = (\sum_{k=1}^{j} MP^k_{i,k1}) \cdot q_1$, equating the tax rate to the marginal value of the external effects at the optimum. With this tax and competitive behavior,

$$\frac{q_i}{q_1} - \frac{t_i}{q_1} = \text{MRS}_{i,1} - \sum_{k=1}^{k} \text{MP}^k_{i,k1} = \frac{p_i}{q_1} = \text{MP}^j_{ji,j1} \quad j = 1, \ldots, J \tag{7.23}$$

or

$$\text{MRS}_{i,1} = \text{MP}^j_{ji,j1} + \sum_{k=1}^{j} \text{MP}^k_{i,k1} \quad j = 1, \ldots, J \tag{7.24}$$

as required for pareto optimality.

A single Pigovian tax is sufficient because the marginal damage to any firm depends only on the aggregate use of factor i. The divergence between the marginal rate of substitution and the marginal external effects, $\text{MRS}_{i,1} - \sum_{k=1}^{J} \text{MP}^k_{i,k1}$ from Eqn (7.21), is independent of j. The only difference from the consumer externality is that the tax equals the value of the marginal external effects rather than the negative of this value, simply because the firm is paying the tax. If the marginal external effect is adverse as in the heat example, the tax is negative (each marginal product MP^k_{ik1} is negative); the producer price p_i must exceed the consumer supply price q_i. Conversely, for external economies, each firm is subsidized in an amount equal to the aggregate marginal external benefit of the activity. With the single tax then, the consumer's marginal rate of substitution is correctly equated to the full social marginal product of factor i in the production of good 1.

Note, finally, that the production model has been written in its most general form. Realistically, any source of pollution affects only a small subset of firms in the economy. In terms of the general model, this simply means that $\text{MP}^k_{ik1} = 0$ for most k in the summation of the external effects.

Three Geometric Interpretations of the Pareto-Optimal Conditions

Three equivalent geometric interpretations have been commonly used in the literature to depict the optimal solution for aggregate production externalities, especially in the context of industrial pollution.

The Market for the Pollutant

The most straightforward representation is in terms of the factor market for i (water), since this is where the external effect actually occurs. In Fig. 7.1, factor demand curve D^{priv} is the horizontal summation of each firm's private demand curve for i, equal to the firm's common private-use value of marginal product between good 1 and water. The supply curve S represents consumer's marginal rate of substitution between i and 1. Without government intervention, the market clears at (X_i^c, p_i^c) with $q_c = \text{MRS}_{i,1} \cdot q_1 = \text{MP}^j_{ji,j1} = , p_i^c$. The curve D^{soc} represents the true social value of marginal product between 1 and i. It differs from D^{priv} at each level of input by a vertical distance equal to the value of the aggregate external marginal damage, $\sum_{k=1}^{J} \text{MP}^k_{i,=k1} \cdot q_1$. The optimum quantity

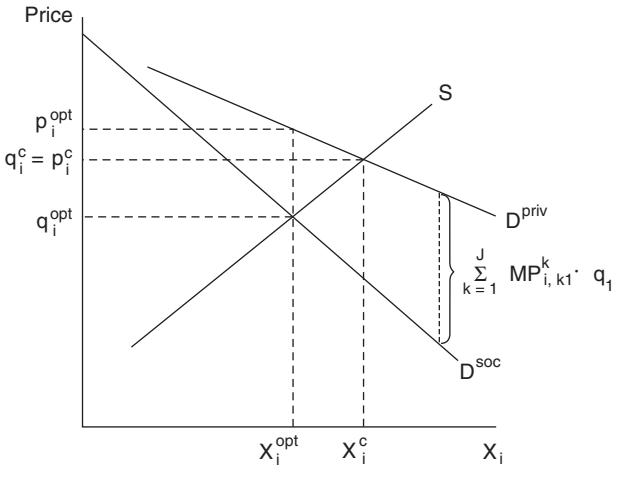

FIGURE 7.1

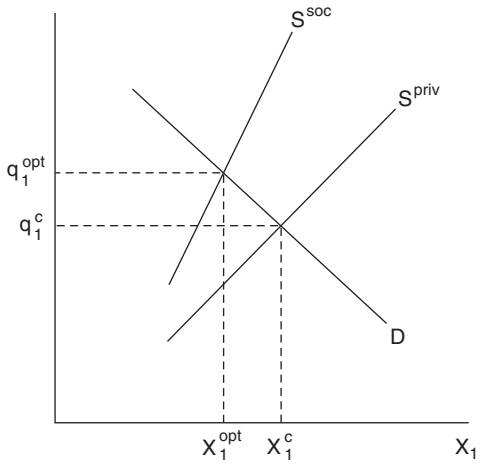

FIGURE 7.2

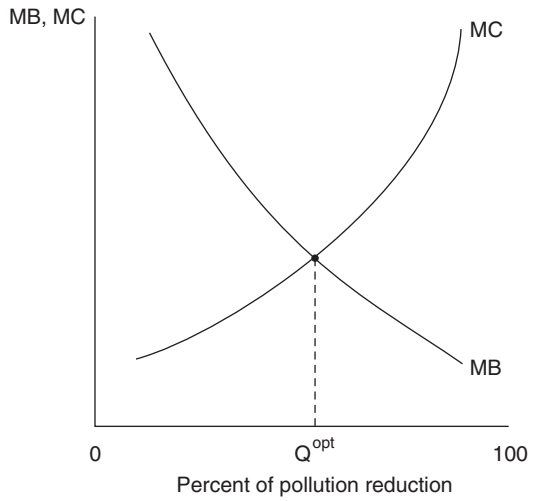

FIGURE 7.3

of i occurs at the intersection of D^{soc} and S, the point at which the social marginal product equals the marginal rate of substitution. If a tax is levied on the use of factor i exactly equal to the aggregate external marginal damage at the optimum X^{opt}, then the decentralized market selects X_i^{opt}, with producer and consumer prices p_i^{opt} and q_i^{opt}, and $q_i^{opt} = p_i^{opt} - t_i^{opt}$.

The Market for Goods That Pollute

An alternative supply—demand interpretation focuses on the market for good 1. Figure 7.2 represents the idea that production of goods generating external diseconomies should be reduced relative to the no-intervention competitive equilibrium, p_i^c. The supply curve S^{priv} is the horizontal summation of each firm's private marginal cost $(qi/\mathrm{MP}^j_{ji,j1})$, the ratio of the price of the input to its marginal product. S^{soc} represents the true social marginal cost of producing good 1, equal to

$$\left(\frac{q_i}{\mathrm{MP}^j_{ji,j1} + \sum_{k=1}^{J} \mathrm{MP}^k_{i,k1}} = \frac{q_i}{\mathrm{MP}^{soc}_{i,1}} \right)$$

Since $\sum_{k=1}^{J} MP^k_{i,=k1} < 0$ for external diseconomies, S^{soc} lies above S^{priv}, as drawn. D is the standard aggregate demand for good 1. In equilibrium, the price q_1 should reflect the social marginal cost of producing good 1, as it does at (X_1^{opt}, q_1^{opt}), and not the private marginal cost, as at (X_1^c, q_1^c). This is equivalent to saying that input prices must equal the value of the social marginal products, not the value of private marginal products.

Extreme care must be taken with this interpretation, however, for two reasons. First, the diagram appears to suggest that a tax on good 1 equal to the divergence between the private and social marginal cost at the optimum X^{opt} can generate a pareto-optimal allocation of resources. This is not true, in general. The Pigovian tax must be on the direct source of the externality to generate the pareto-optimal conditions, in this case on the use of factor i. The purpose of the tax is to change the firms' incentive to use water. Any output effects from the tax on water happen indirectly as the result of increasing the marginal cost to the firms of using water. Second, output effects in the presence of externalities are not as straightforward as this partial equilibrium diagram might suggest. William Baumol and Wallace Oates have demonstrated that with combined production and consumption externalities, which may well exist with industrial pollution, the conditions required to guarantee output reductions for activities that generate external diseconomies are fairly restrictive.[4]

The Optimal Reduction in Pollution

A final geometric interpretation, especially common in pollution analysis, says that the external damage should be reduced until the marginal benefit just equals the marginal cost of the reduction. In Fig. 7.3, Q^{opt} represents the optimal amount of external damage. The diagram is a useful device for showing that, in general, zero damage (zero pollution) is typically not the pareto-optimal solution.

Figure 7.3 can be directly related to Fig. 7.1 in the following manner. The marginal benefit of reducing external damage is the negative of the marginal cost of increasing the external damage. In Fig. 7.1, this marginal

4. Baumol and Oates (1975, Chapter 7). For a similar comprehensive analysis with consumer externalities, see Diamond and Mirrlees (1973); also, Sadka (1978). The earliest recognition of possible output anomalies with externalities is generally credited to Buchanan and Kafolgis (1963).

cost is $\sum_{k=1}^{J} \mathrm{MP}_{i,=k1}^{k} \cdot q_1$, the value of the reduction in output of good 1 through the externality caused by a marginal increase in factor i. Therefore, the marginal benefit curve of Fig. 7.3 equals the vertical distance between D^{priv} and D^{soc} in Fig. 7.1. The marginal cost of reducing damage is an opportunity cost. It equals, at each quantity of factor input, the marginal private-use value of factor i in production of good 1 (D^{priv} in Fig. 7.1), less the value at which consumers are willing the supply factor i (curve S in Fig. 7.1). Therefore, the marginal cost curve in Fig. 7.3 equals the vertical distance between curves D^{priv} and S in Fig. 7.1. Q^{opt} in Fig. 7.3 thus corresponds to X_i^{opt} in Fig. 7.1: MB = MC in terms of external damage reduction when the distance between D^{priv} and D^{soc} equals the distance between D^{priv} and S in the market for factor i.

Internalizing the Externality

Correcting for an externality does not necessarily require a Pigovian tax. There could be other options.

Suppose that a single conglomerate owned all the firms affected by a particular externality. In terms of our general model, this would include every single firm in the economy, but externalities will be much less pervasive in actual cases. If one firm does own all affected firms, then its desire to maximize profits gives it the proper incentive to account for the externality. The government need not intervene because the firm effectively takes on the role of the omniscient social planner.

Our model may be unduly general, but it can be used to illustrate this point quite effectively. The single firm would solve the following problem: Allocate the goods and factors among all production sites to maximize joint profits. Formally,[5]

$$\underset{(X_{kn})}{max} \left(\sum_{k=1}^{J} \sum_{n=1}^{N} p_n X_{kn} \right)$$
$$\text{s.t.} f^k \left(X_{kn}; \sum_{j=1}^{J} X_{ji} \right) = 0$$

with the corresponding Lagrangian:

$$\underset{(X_{kn})}{maxL} = \sum_{k=1}^{J} \sum_{n=1}^{N} p_n X_{kn} + \sum_{k=1}^{J} \lambda^k f^k \left(X_{kn}; \sum_{j=1}^{J} X_{ji} \right)$$

The first-order conditions for this problem are

$$X_{kn}: p_n = -\lambda^k f_{kn}^k \quad n \neq i, k = 1, \ldots, J \tag{7.25}$$

$$X_{ki}: p_i = -\lambda^k f_{ki}^k - \sum_{j=1}^{J} \lambda^j f_i^j \quad k = 1, \ldots, J \tag{7.26}$$

5. Here, $n = 1, \ldots, N$ and includes i.

Expressing the pareto-optimal conditions in terms of good 1 yields

$$\frac{p_n}{p_1} = \frac{f_{kn}^k}{f_{k1}^k} \quad n \neq i; \; k = 1, \ldots, J \tag{7.27}$$

$$\frac{p_i}{p_1} = \left(\frac{f_{ki}^k}{f_{k1}^k} \right) + \sum_{k=1}^{J} \left(\frac{f_i^k}{1} \right) \quad k = 1, \cdots, J \tag{7.28}$$

If all markets are perfectly competitive and there are no taxes, then

$$q_n = p_n \; n = 1, \ldots, N$$

Thus, combining utility and profit maximization,

$$\frac{q_n}{q_1} = \frac{U_n}{U_1} = \frac{f_{kn}^k}{f_{k1}^k} \frac{p_n}{p_1} \quad n, \neq i; \; k = 1, \ldots, J \tag{7.29}$$

and

$$\frac{q_i}{q_1} = \frac{U_i}{U_1} = \left(\frac{f_{ki}^k}{f_{k1}^k} \right) + \sum_{k=1}^{N} \left(\frac{f_{ki}^k}{f_{k1}^k} \right) = \frac{P_i}{P_1} \quad k = 1, \ldots, J \tag{7.30}$$

the required pareto-optimal conditions.

This example illustrates two important points. The first relates to modeling strategy. Any situation involving only production externalities does not require a full general equilibrium model to determine the pareto-optimal conditions. All one need assume is that society is trying to maximize total profits in the economy at fixed producer prices, subject to all the production constraints. A number of researchers have exploited this property and ignored the demand side entirely. The only caveat is that the optimal prices, p_n, cannot be determined without specifying consumer preferences as well. Hence, all profit-maximizing specifications implicitly assume that the prices in the objective profit function are set equal to their values at the full pareto optimum.

The second point is that some decision-making unit has to internalize an externality in order to achieve pareto optimality. This is a fundamental prerequisite for any solution to a technological externality, whatever form the externality may take.

One possibility is a bargaining solution among the affected firms. The nature of the bargain is cartel-like. In our example, the firms agree to adjust the production of the externality-generating activity to maximize group profits and then further agree on how to divide the increased profits among themselves. This is the solution envisioned by Coase in his famous theorem. Firms certainly have an incentive to internalize the externality in this way because it is potentially a pareto-superior outcome. We will see in Chapter 20, however, that private information about the externality can undermine the incentive to bargain efficiently.

A number of practical problems remain for Coase-style bargaining even with perfect information. One is that a bargaining solution is undoubtedly infeasible if large numbers of firms are affected by the externality. The second concerns the nature of their bargain. The bargaining solution requires that the firms behave in cartel-like fashion in accounting for the externality, but they cannot also use their new-found monopoly power to raise prices to consumers. The firms must remain price takers, or some of the first-order conditions, Eqn (7.27), will not hold. Finally, the bargaining envisioned by Coase requires collusion by the firms and may run afoul of the US antitrust laws.

If the firms cannot or will not internalize the externalities by themselves, then the government must force society to "see" the correct pattern of interrelationships by setting Pigovian taxes (or subsidies). In practice, however, effective internalization by the government sector may also be difficult to achieve. This is especially likely with a federalist system of national, state, and local governments. As will be discussed in detail in Part V, one of the main theoretical problems with a federalist system of governments is that the jurisdictional boundaries of any one government seldom correspond to the pattern of externalities present in the economy. This is particularly true for most forms of air and water pollution. Individual state and local governments often cannot internalize all the external effects simply because many of the affected citizens or firms are not located within their jurisdictions. The national government could theoretically internalize all externalities, but it seldom has the flexibility to offer variable policy solutions tailored to specific local pockets of external effects, especially if the externalities cut across lower level jurisdictions. This jurisdictional dilemma may well go a long way toward explaining why the United States has never been able to mount a very effective antipollution policy.

Additional Policy Considerations

A number of additional policy considerations can best be analyzed in the context of a simpler model in which only one firm is the source of the externality. This may actually be a more realistic model for many externalities, such as a single-source industrial polluter.

Suppose firm 1 produces a product or by-product, call it z_1, that enters the production function of all other firms in the economy but is a decision variable only for the first firm. z_1 has no effect on consumers. Assume, further, that all other goods and factors X are purely private, and that all firms are price takers operating in competitive markets. An example might be a firm situated on a river and engaging in some polluting activity that affects all other firms located downstream from it.

In this model, the production functions can be represented as

$$f^1\left(X_{1n}; z_1\right) = 0 \tag{7.31}$$

$$f^k\left(X_{kn}; z_1\right) = 0 \quad n = 1, \ldots, N; \; k = 2, \ldots, J \tag{7.32}$$

The government's problem is

$$\begin{aligned} \max_{(X_{\ln}; X_{kn}; z_1)} \; & U\left(\sum_{j=1}^{N} X_{jn}\right) \\ \text{s.t.} \; & f^1\left(X_{\ln}; z_1\right) = 0 \\ & f^k\left(X_{kn}; z_1\right) = 0 \end{aligned}$$

with the corresponding Lagrangian

$$\begin{aligned} \max_{X_{\ln}; X_{kn}; z_1} \; L = {} & U\left(\sum_{j=1}^{N} X_{jn}\right) + \lambda^1 f^1(X_{\ln}; z_1) \\ & + \sum_{k=2}^{J} \lambda^k f^k(X_{kn}; z_1) \\ & k = 2, \ldots, J; \; n = 1, \ldots, N \end{aligned}$$

The first-order conditions for this problem are

$$U_n = \lambda^k f^k_{kn} = \lambda^1 f^1_{1n} \quad n = 1, \ldots, N; \; k = 2, \ldots, J \tag{7.33}$$

$$\lambda^1 f^1_{z_1} + \sum_{k=2}^{J} \lambda^k f^k_{z_1} = 0 \tag{7.34}$$

Expressing the pareto-optimal conditions in terms of good 1 yields

$$\frac{U_n}{U_1} = \frac{f^k_{kn}}{f^k_{k1}} = \frac{f^1_{1n}}{f^1_{11}} \quad n = 1, \ldots, N; \; k = 2, \ldots, J \tag{7.35}$$

and

$$\frac{f^1_{z1}}{f^1_{11}} + \sum_{k=2}^{J} \left(\frac{f^k_{z1}}{f^k_{k1}}\right) = 0 \tag{7.36}$$

Equation (7.36) follows from the consumer's indifference to which firms supply goods or buy factors $\lambda^j f^j_{j1} = \lambda^1 f^1_{l1} = U_1 \neq 0, j = 2, \ldots, J$. This assumption can then be used to remove the Lagrangian multipliers from Eqn (7.34) by selective substitution in the denominators, as demonstrated in the aggregate externality case. A number of important policy implications follow from the first-order conditions.

Taxing the Externality

The government can achieve the pareto-optimal conditions by setting a unit tax on firm 1's production of z_1, such that

$$t_z = -\sum_{k=2}^{J} \left(\frac{f^k_{z1}}{f^k_{k1}}\right) \cdot q_1 \tag{7.37}$$

equal to the value of the aggregate marginal external effect from producing z_1. All other goods and factors are untaxed.

This is the standard Pigovian tax; it achieves the pareto-optimal conditions because the firm sets $q_1 \cdot \left(\frac{f^1_{z1}}{f^1_{11}}\right) = t_z$.

Taxing and Subsidizing Everything Else

A unit tax (subsidy) on the externality-generating activity works by changing the vector of relative prices in the economy from their values in the no-intervention competitive situation to the values necessary to support the pareto optimum. Only relative prices determine the allocation of resources. This implies that any set of absolute prices that maintains the unique vector of pareto-optimal relative prices is an admissible solution to the externality problem. An infinity of absolute prices satisfy the optimal relative price vector, including a vector of prices in which the externality-generating activity is not taxed. An interesting problem, then, is to find the set of taxes (and subsidies) that generates the pareto-optimal allocation given that, for some reason, the externality-generating activity cannot be taxed.

The following set of taxes on firm 1 achieves the pareto optimum:

$$\begin{aligned} t^1_1 &= a = \left(\frac{f^1_{11}}{f^1_{z1}}\right) \cdot \sum_{k=2}^{J} \left(\frac{f^k_{z1}}{f^k_{k1}}\right) q_1 \\ t^1_n &= a\left(\frac{f^1_{1n}}{f^1_{11}}\right) \quad n = 1, \ldots, N \\ t^1_z &= 0 \end{aligned} \tag{7.38}$$

The tax on good 1, a, equals the marginal increase in z_1 resulting from a unit increase in good 1 by firm 1 (first term), times the marginal decrease in good 1 across all firms per unit increase in z_1. The tax on one of the private goods n equals the marginal increase in good 1 by firm 1 per unit increase in good (factor) n, multiplied by the aggregate marginal external effect of an increase in good 1, given by a. As such, the two taxes account for the aggregate external effects of all firm 1's activities except its production of z_1. In other words, the taxes indirectly account for the externality caused by firm 1.

To see that these taxes generate the pareto-optimal conditions, consider firm 1's use of any good or factor n and good 1. Firm 1 equates

$$q_n + t^1_n = \frac{f^1_{1n}}{f^1_{11}}(q_1 + a) \tag{7.39}$$

or

$$q_n = \frac{f^1_{1n}}{f^1_{11}} q_1 \tag{7.40}$$

as required for condition (7.35).

Next, consider the firm's use of z_1 and good 1. The firm equates

$$(q_1 + a)\frac{f^1_{z1}}{f^1_{11}} = t_z = 0 \tag{7.41}$$

as required for condition (7.36).

Notice that the government should not levy any taxes on firms $k = 2, \ldots, J$. Since z_1 is not a decision variable for firms $k = 2, \ldots, J$, their production of the other goods and factors can be left untaxed.

This exercise emphasizes the importance of taxing the source of the externality if possible, a point mentioned above in the discussion of Fig. 7.2. Otherwise, the government must tax (or subsidize) all goods and factors that directly substitute for the externality-causing activity in production, and that can be a very large number.

Subsidizing or Compensating the Victims

The tax analysis also shows that the government cannot merely subsidize (tax) firms $k = 2, \ldots, J$ for the damage (gain) caused by firm 1's production of z_1. No matter what form the subsidy (tax) may take, society cannot possibly satisfy the pareto-optimal condition, Eqn (7.36), if firm 1 is not taxed appropriately, Firm 1, if untaxed, will produce z_1 until $f^1_{z_1}/f^1_{11} = 0$, contrary to the requirements of pareto optimality. Furthermore, if the government chooses to subsidize the other firms by means of a unit subsidy (tax) on any of the other firm's outputs or inputs or any other type of subsidy that changes their first-order profit-maximizing conditions, then a subset of condition (7.35) must fail as well. Firms $k = 2, \ldots, J$ and firm 1 would face different prices for at least one of the N goods and factors. Consequently,

$$\frac{f^k_{kn}}{f^k_{k1}} \neq \frac{f^1_{1n}}{f^1_{11}} \tag{7.42}$$

for some good or factor n and some firm k, contrary to the pareto-optimal conditions.

Furthermore, suppose the government chooses to tax firm 1's use of z_1 and does so correctly (assume the externality is a diseconomy). The government can use the tax revenues to compensate some or all of the remaining firms $k = 2, \ldots, J$ (the "victims" of the externality), but it must do so in a lump-sum fashion, z_1 is a lump-sum event from the point of view of the other firms, and the subsidy must be, too. Otherwise some of the pareto-optimal conditions, Eqn (7.35), will fail to hold.

Partial Taxes and Subsidies

Regarding the policy of taxing z_1, the government need not place a unit tax on the entire production of z_1. It can instead

tax the production of z_1 only above some arbitrary minimal level $\bar{z}_1$, perhaps, using the pollution example again, a level judged to be harmless. Alternatively, it can subsidize firm 1 for reducing z_1 below some other arbitrary level, $\bar{\bar{z}}_1$, perhaps the level of z_1 at the untaxed, prepolicy, competitive equilibrium. The objective profit function for firm 1 with each of these alternatives is

$$\text{Option a: } \sum_{n=1}^{N} p_n X_{1n} - t_z z_1 \quad (\text{tax entire } z_1) \tag{7.43}$$

$$\text{Option b: } \sum_{n=1}^{N} p_n X_{1n} - t_z(z_1 - \bar{\bar{z}}_1) \quad (\text{tax } z_1 \text{ above } \bar{\bar{z}}_1) \tag{7.44}$$

$$\text{Option c: } \sum_{n=1}^{N} p_n X_{1n} + s_1(\bar{\bar{Z}}_1 - Z_1) \quad (\text{subsidize reduction of } z_1 \text{ below } \bar{\bar{z}}_1) \tag{7.45}$$

where $s_z =$ a unit subsidy.

These profit functions all lead to the same first-order conditions if firm 1 maximizes any one of them subject to its production constraint $f^1(X_{1n};\ z_1) = 0$. We know that profit function (7.43) generates the proper pareto-optimal conditions with

$$t_z = -\sum_{k=2}^{J}\left(\frac{f^k_{z1}}{f^k_{k1}}\right) \tag{7.46}$$

Therefore, so too must Eqns (7.45) and (7.46) so long as

$$s_z = t_z = -\sum_{k=2}^{J}\left(\frac{f^k_{z1}}{f^k_{k1}}\right) \tag{7.47}$$

$\bar{z}_1, \bar{\bar{z}}_1, t_z$, and s_z in Eqns (7.44) and (7.45) are all parameters fixed by the government. Thus, the terms $t_z \cdot \bar{z}_1$ and $s_z \cdot \bar{\bar{z}}_1$ in Eqns (7.44) and (7.45) cannot affect the first-order conditions for profit maximization.

Entry, Exit, and Optimality in the Long Run

Policy options b and c may cause problems in the long run if the government is not careful, however, a point first demonstrated by Baumol and Oates in their book *The Theory of Environmental Policy* (Baumol and Oates, 1975, Chapter 12). Consider the subsidy, option c. In the unlikely event that $\bar{\bar{z}}_1$ happens to equal the value of z_1 at the full pareto optimum, firm 1 receives no net subsidy and no problem arises. One would expect $\bar{\bar{z}}_1$ to be set at a value greater than z_1^{opt}, however, in which case firm 1 actually receives a subsidy. If so, and the economy was at a zero-profit competitive equilibrium before the subsidy, other firms now have an incentive to enter the industry represented by firm 1 to receive the same subsidy. In effect, policy option c raises the marginal costs of firm 1 by an amount related to the unit subsidy s_z, while simultaneously lowering its average costs because of the lump-sum subsidy, $s_z\bar{\bar{z}}_1$. The average cost-lowering effect occurs so long as $\bar{\bar{z}}_1 > z_1^{\text{opt}}$. The situation is depicted in Fig. 7.4. To interpret the diagram, think of good 1 as the output of firm 1 and that a large number of such firms comprise industry 1. The original no tax (subsidy) long-run equilibrium for each firm in the industry is at (X_0^1, q_0).

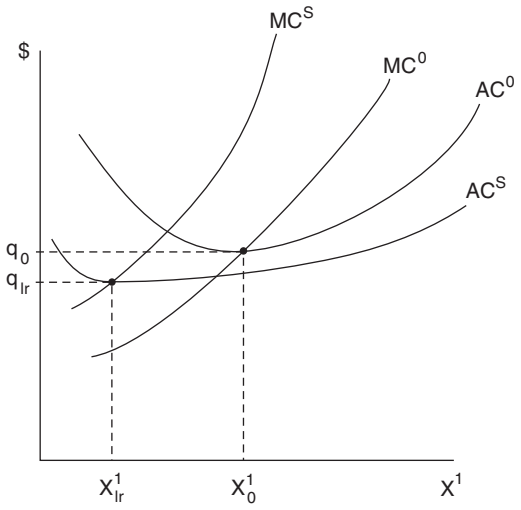

FIGURE 7.4

With policy option c, each firm's marginal costs shift upward from MC^0 to MC^S, as required for optimality. But their average costs fall from AC^0 to AC^S because, net, they are subsidized by amount $s_z \cdot (\bar{\bar{z}}_1 - z_1)$. The new long-run equilibrium is at (X_{1r}^1, q_{1r}). Although each firm's production has decreased from X_0^1 to X_{1r}^1, entry of new firms leads to the lower price and an increase in industry output. As a result, total production of z_1 may actually rise, surely an unwanted result.

The problem is that a subsidy given to producers in industry 1 is not truly a lump-sum subsidy for the economy as a whole in the long run, if other firms have the option of entering the industry. As discussed in Chapter 2, a lump-sum subsidy has the property that economic decisions cannot alter the size of the subsidy. Thus, to make these subsidies truly lump sum in the long run, they must either be offered to all firms whether or not they actually enter the first industry or be given only to the original firms in the industry and not to new entrants. Because governments are probably not going to do either of these, the safest policy is simply a unit tax on the full amount of z_1, policy option a.

Strictly speaking, the Baumol–Oates subsidy problem cannot arise in the model as presented above because of our implicit assumption that only firm 1 can produce z_1. Hence, the other firms $k = 2, \ldots, J$ are not even potential entrants into industry 1. One can imagine a different model,

however, in which firms $k = 2,\ldots, J$ can produce z_1 in the long run but choose not to without a subsidy, given the going market prices $(p_1,\ldots, p_N)$, and the form of the $(J-1)$ production functions $f^k(\ )$. This is the type of model Baumol and Oates have in mind.

Policy option b also fails if the alternative Baumol–Oates model really applies in the long run. It raises average costs so long as $z_1^{\text{opt}} > \overline{Z}_1$ but not by the same amount as a tax on the full amount of z_1. If type 1 firms can become other kinds of firms, not enough of them will exit industry 1 in the long run. In effect, the term $t_z \cdot \overline{Z}_1$ acts as a locational subsidy and is not consistent with pareto optimality.

The original conclusion stands: The safest policy is a straight unit tax of the full amount of z_1. With this policy it does not matter which of the two models actually applies. It is always pareto optimal.

Bargaining in the Long Run with Entry and Exit

Free entry has troubling implications for Coase-style bargaining solutions to externalities. It turns out to make efficient bargaining extremely problematic. Earlier we showed that joint profit maximization among all the firms associated with the externality, both the generators or receivers of the externality, satisfies the first-best pareto-optimal conditions, as Coase had surmised in his theorem. The efficient bargain rests on four assumptions:

1. The property rights to control the extent of the externality and the disposition of the profits are assigned to some decision maker.
2. Prices are taken as given.
3. Bargaining among the firms is costless.
4. The number of firms in each industry is fixed.

The fourth assumption is crucial to the Coasian efficiency result. The possibility of entry into the externality-generating or externality-receiving industries adds a new dimension to the bargain process that severely limits the chances for an efficient solution. Unfortunately, the assumption of free entry in the long run goes hand in hand with the assumption of competitive, price-taking behavior, which is also necessary for efficient bargains.

Jonathan Hamilton, Eytan Sheshinski, and Steven Slutsky (HSS) explored the problems of bargaining in a general equilibrium model with a production externality that is about as simple as such a model can be (Hamilton et al., 1989). Their model consists of just two goods, X and Y, with the production of X conferring an aggregate external diseconomy in the form of pollution on Y. Labor is the only factor of production, and the consumers' utility is additively separable in labor to remove income effects from the model. Producers operate in competitive markets with free entry (exit) in the long run. They take prices as fixed.

HSS consider three types of property rights that might be associated with the externality: a liability rule, a complete property right, and an ultracomplete property right. Under a liability rule, agents are assigned property rights only by entering an industry. The rule might take the form of a right of X producers to collect bribes from Y producers for reducing pollution in the X industry, or a right of Y producers to collect damages resulting from the externality. A complete property right exogenously assigns the rights to fees or compensation associated with the externality, which can then be purchased from the owner. A complete property right would give the owner the right to determine the amount of pollution. Ownership of the property right is independent of entry into one of the industries. An ultracomplete property right extends the complete property right by also granting the owner control over entry in both the industries. The owner can collect entry fees from firms in either industry.

A liability rule leads to inefficient bargains with entry for the same reason that partial taxes and subsidies do. The number of firms in one or both industries is nonoptimal. For example, a liability rule encourages too much entry into the X industry if X producers have the right to collect bribes or too much entry into the Y industry if Y producers have the right to collect damages.

Assignment of complete property rights is also incompatible with the efficient solution in the long run. Consider the right to control the amount of pollution in the X industry, which in the HSS model is the same as controlling the total output of X. Efficiency requires that the Y industry reach its zero-profit equilibrium without interference of any kind (the complete property right cannot be an exogenous right to damages in the Y industry). The zero-profit equilibrium also implies that the property right owners cannot extract any income from the Y producers, such as through bribes. Therefore, the owners' incentive is to ignore the Y industry and maximize profits in the X industry at the fixed competitive price P_x. But, ignoring the Y industry ignores the external damage caused by the production of X, so that the profit maximizing solution cannot be the efficient solution.

Assignment of ultracomplete property rights also cannot sustain an efficient equilibrium with positive production of X and Y under the assumptions of price-taking behavior, costless bargaining among firms, and free entry. The problem is that the externality causes the second-order conditions to fail at the efficient allocation.

The essence of the failure here concerns the relation of the fixed prices to the minimum long-run average costs in each industry (LRAC_{min}). Industry X has a unique LRAC_{min}, say at P_x^{min}. The LRAC_{min} in industry Y depends on the value of X; the lowest LRAC_{min} occurs at $X = 0$, say

at a value $P_y^{\min}$. Suppose the fixed competitive prices are the minimum possible values $P_x^{\min}$ and $P_y^{\min}$. Then, one of the following is true:

1. $X = 0$ and there are zero profits in the Y industry.
2. $X > 0$, and $Y = 0$ (since $P_y^{\min} < \text{LRAC}_{\min}$ in the Y industry with $X > 0$); also, there are zero profits in the X industry (since $P_x^{\min} = \text{LRAC}_{\min}$).

In either case the property rights owner earns zero profit.

Suppose that the first-best efficient equilibrium is an interior solution with $X,Y > 0$. Then $P_x > P_x^{\min}$ and $P_y > P_x^{\min}$. In addition, profits in the Y industry must be zero at the efficient equilibrium. HSS show that the efficient solution fails the second-order conditions for maximizing the fees collected by the property rights owner. The solution is a saddle point, with profit (fee) maximizing in Y (at zero profit) but profit (fee) minimizing in X. Hence, the efficient equilibrium is not sustainable in the long run even with ultracomplete property rights.

The intuition as to why X is a profit minimum is as follows:

1. An increase in X increases profits in the X industry at fixed P_x and $P_x > \text{LRAC}_{\min}$.[6] The efficient solution reduces the profits in X below the maximum possible profit because it accounts for the externality on the Y producers. True, the increase in X increases costs in Y and drives Y producers out of business at the fixed P_y. But the property rights owner earns zero profits from the Y industry anyway at the efficient equilibrium, so no fees are lost from the Y industry.
2. A reduction in X reduces the profits in X. But it also lowers the costs of producing Y and leads to profits in the Y industry at the fixed P_y. It turns out that the profit-increasing effect in the Y industry dominates in the HSS model.

HSS show that costless Coase bargaining can be efficient in their model, but only if the owners of the property rights, whether complete or ultracomplete, can engage in costless, all-or-none bargains with the consumers as well as the firms. The owners bargain to keep prices at the social optimum and take from the consumers all their utility (consumer surplus) above the utility they would receive if the owners behaved as monopolists. In other words, the owners engage in first-degree price discrimination. The ability to capture the extra consumer surplus provides the owners with the incentive to produce at the social optimum since it maximizes total consumer surplus.

6. The assumption of fixed prices is crucial for efficiency. If the owners of the ultracomplete property rights see the demand curves of the consumers and can manipulate prices, they will act as profit-maximizing monopolists, which certainly cannot yield the efficient solution.

This is hardly the decentralized bargaining that Coase envisioned, however. To the contrary, the knowledge and market power of the property rights owners would have to be equivalent to that of an omniscient socialist planner. The conclusion to be drawn from HSS's analysis is clear: Efficient decentralized bargaining solutions to externalities in a competitive market environment are patently unrealistic.[7] This is discouraging, the more so because the government's ability to design optimal Pigovian taxes is also highly problematic. Externalities pose difficult problems indeed for market economies.

Bargaining Costs versus Property Rights

Dan Usher offers an appropriate concluding general perspective on Coasian bargains (Usher, 1998). In his view, the key assumption behind the Coase theorem is that bargaining is costless and not the assignment of property rights. If bargaining were truly costless, then economic agents would naturally come together and make whatever arrangements were required to reach a mutually advantageous pareto optimum. The assignment of property rights would be irrelevant. In terms of the HSS model, consumers would join with firms and the owners of the property rights in the all-or-none bargains required for economic efficiency. Indeed one can imagine economic agents worldwide bargaining to maximize, and divide, total world income. The Coase theorem is a tautology under costless bargaining.

The truth is that bargaining is almost always costly and generally the more so as the number of agents in the bargain increases. This explains why societies have chosen markets and command systems to allocate resources rather than relying exclusively on bargaining. Under costly bargaining, the assignment of property rights mostly determines who gets to join the bargaining process. It does not necessarily determine whether the bargains will be (second-best) efficient. Governments establish property rights primarily because they agree to enforce contracts, and enforcement is easier if property rights have been assigned.

7. HSS also consider the combination of bargaining with Pigovian taxation. They imagine that the government attempts to set optimal Pigovian taxes under the assumption that the producers in each industry are independently maximizing profits. In fact, bargaining is occurring to maximize the income of the property rights owners and the government is unaware of this. This policy environment is second best because the behind-the-scenes bargaining is private information from the government's perspective. Not surprisingly, the Pigovian tax is no longer efficient in the presence of bargaining. Efficiency requires a highly complex and nonlinear tax scheme even in the simple HSS model. Moreover, the tax revenues must be returned lump sum to the producers to support the efficient solution. The Pigovian tax revenues are returned lump sum to the consumers in the first-best policy environment.

We will return to Coasian bargains one last time in Chapter 20, when we consider the effects of imperfect information on bargaining outcomes. Imperfect information can lead to inefficient bargains even when only two agents are bargaining and the bargaining process is otherwise costless.[8]

CONCLUDING COMMENTS: THE PROBLEM OF NONCONVEX PRODUCTION POSSIBILITIES

The analysis in this chapter has assumed that aggregate production possibilities are strictly convex. This is a crucial assumption, for without it the tax policies and their equivalents offered as a means of achieving pareto optimality may be only locally optimal. They may not represent a global optimum. Unfortunately, production externalities themselves can generate significant nonconvexities, so that the assumption may not be valid.[9]

Analyzing the special problems caused by nonconvex production possibilities (increasing returns, decreasing costs) is premature, as the general theoretical treatment of nonconvexities appears in Chapter 9. But the crux of the problem can be seen with reference to a simple two-good, one-factor economy.

Suppose that goods X_1 and X_2 are produced with linear technology by a single factor of production, L (labor). If there are no externalities, then the production-possibilities frontier is a straight line, AB, reflecting constant opportunity costs, as depicted in Fig. 7.5. If X_1 generates an external diseconomy for X_2, however, then the quantity of X_2 must lie below AB at each X_1, except at the end points. Hence, assuming the frontier is continuous, it must contain a nonconvex segment near the end point B, as depicted in Fig. 7.6.

To see the potential local–global problem, suppose society initially ignores the externality, thereby underestimating the true costs of producing X_1, and achieves an equilibrium at D on indifference curve I_0 in Fig. 7.6, on the nonconvex region of the frontier. Opportunity costs are incorrectly measured by the slope of I_0 at D. A Pigovian tax reduces production of X_1 and moves society to point T, where indifference curve I_1 is tangent to the frontier. Although T is an improvement over D, it is only a local optimum. The global optimum is at point G, the tangency of I_3 with the frontier, and a Pigovian tax cannot possibly

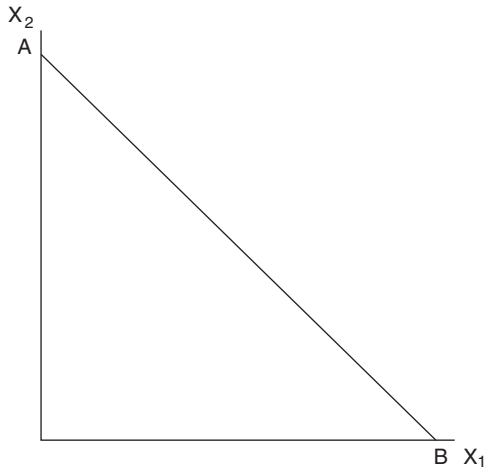

FIGURE 7.5

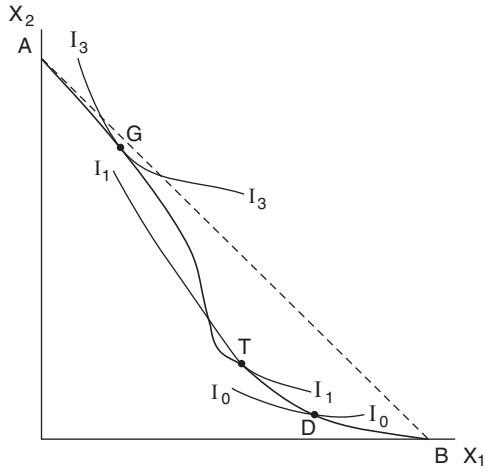

FIGURE 7.6

achieve G starting from D. A similar demonstration applies to the case of external economies.

A common example is a laundry (X_2) located downwind from a factory (X_1). A Pigovian tax on the smoke emitted by the factory may benefit the laundry by reducing the smoke pollution. The least cost solution, however, may simply be to have the laundry move upwind from the factory and thereby avoid all (or almost all) of the smoke pollution.

The laundry–factory example is a specific instance of a more general question: How must optimal policies be adjusted if the victims of an external diseconomy such as pollution can partially or completely defend themselves from the external effects? The example illustrates one wrinkle, that defensive strategies can themselves give rise to nonconvexities. We will consider the question of defensive expenditures in more detail in Chapter 8. Waste treatment, a defensive strategy, is an important part of the

8. Coasian bargaining has fared well in experimental settings in which the subjects are well informed and bargaining is relatively costless. As one example of bargaining in the presence of an aggregate externality, consult Harrison et al. (1987).

9. See Baumol and Oates (1975, Chapter 8) for an excellent detailed analysis of the nonconvexity issue and the important distinction between local and global solutions to externality problems.

United States' fight against pollution, and it is justified on the basis of decreasing cost (nonconvex) production.

REFERENCES

Baumol, W.J., Oates, W., 1975. The Theory of Environmental Policy. Prentice-Hall, Englewood Cliffs, N.J.

Buchanan, J., Kafolgis, M., June 1963. A note on public goods supply. American Economic Review Vol. 53 (3), 403–413.

Diamond, P., Mirrlees, J., February 1973. Aggregate production with consumer externalities. Quarterly Journal of Economics Vol. 87 (1), 1–24.

Hamilton, J., Sheshinski, E., Slutsky, S., July 1989. Production externalities and long-run equilibria: bargaining and pigovian taxation. Economic Inquiry Vol. 27 (3), 453–472.

Harrison, G., Hoffman, E., Rutstrom, E., Spitzer, M., June 1987. Coasian solutions to the externality problem in experimental markets. Economic Journal Vol. 97, 388–402.

Sadka, E., February 1978. A note on aggregate production with consumer externalities. Journal of Public Economics Vol. 9 (1), 101–105.

Usher, D., October 1998. The Coase theorem is tautological, incoherent, or wrong. Economic Letters Vol. 61 (1), 3–11.

Chapter 8

An Application of Externality Theory: Global Warming

Chapter Outline

The externality that has been on everyone's mind since the 1990s is global warming. Climate scientists universally agree that average temperatures are rising worldwide. There is also an overwhelming consensus among them that the emission of greenhouse gases (GHGs) is a major contributor to the warming trend. The GHGs include carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), and sulfur hexafluoride (SF_6). These gases build up in the atmosphere and trap heat radiating from the earth's surface, hence the name greenhouse gases. Among the GHGs, carbon dioxide is targeted as the one to control to halt global warming. It does not cause the most heat retention per ton of emission; for example, methane has 21 times the heat retention properties of CO_2 100 years out. But since CO_2 is emitted from the burning of fossil fuels, emissions of CO_2 are many times larger than the emissions of the other gases.[1]

Assuming the scientific consensus is correct, CO_2 emissions from the burning of fossil fuels is an example of a policy-relevant externality because global warming generates a large number of direct third-party effects that are not accounted for in the market economy. Moreover, it is an externality with some properties that were not addressed in Chapters 6 and 7. Chapter 8 completes our presentation of externalities by analyzing these properties, many of which apply to other externalities as well. We begin with an analysis of consumption–production externalities since global warming is a leading example of that category of externality.

CONSUMPTION–PRODUCTION EXTERNALITIES

A policy-relevant, technological, consumption–production externality is an externality in which some production (consumption) activity enters the utility function (production function) of a least one consumer (producer). The externality may affect other producers (consumers) as well.

The CO_2 emissions from the burning of fossil fuels by households and businesses are a clear example. Regarding production, global warming has potentially enormous effects on agricultural production throughout the world, both positively and negatively. Some regions that currently

1. Concern is building about emissions of methane gas as fracking to extract oil and natural gas reserves trapped in rock has grown. The fracking process releases methane gas, which can escape into the atmosphere if it is not adequately contained on site. The US Environmental Protection Agency Web site has a wealth of information on the climate effects of the GHGs as well as on other pollutants. The methane/CO_2 comparison is from epa.gov/climatechange/science/indicators/ghg/index.html.

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00008-X

produce a valued crop such as corn may become too hot and dry to grow corn if global warming continues, whereas other regions that are now too cold to grow corn may become warm enough to support the crop. Households are affected as well. Many seaside homes may have to be abandoned if the level of the oceans continues to rise, and coastal cities may have to erect extensive dikes to protect their residents and businesses. Worse yet, many people who live in poor countries with relatively warm climates that struggle to provide enough food for their citizens may starve to death if agricultural production in these countries declines. Conversely, people living in frigid climates may enjoy much more pleasant, temperate climates 50 years from now.

Unfortunately, an analysis of consumption–production externalities requires the full general equilibrium model of Chapter 2 or its equivalent to capture the extent of the external effects. Condensing the model as we did for consumption and production externalities would hide essential features of the externality. The notational requirements alone are formidable in a complete general equilibrium model. But, having worked through the general equilibrium model of Chapter 2 and the pure consumption and production externality models of Chapters 6 and 7, the analysis of consumption–production externalities is reasonably straightforward and predictable.

Consider the general case of an aggregate consumption–production externality, in which the aggregate use of some factor in production (e.g., fuel) enters into every person's utility function and every firm's production function through the externality it generates, with all other goods and factors purely private. This is the most extensive possible example of an aggregate consumption–production externality and reasonably appropriate for the analysis of global warming. As it happens, emissions of any of the GHGs from any source mix completely into the atmosphere, such that only the aggregate emissions matter for global warming. The particular sources and locations of the emissions are irrelevant. In addition, the extent of the externality is truly global as the name global warming implies, because the warming effect of the GHGs reaches all parts of the globe. This property makes global warming a difficult problem to solve since it requires the cooperation of all the major emitting countries, a problematic outcome in almost all matters of world affairs. Later on in the chapter, we will consider some geopolitical factors that play a role in choosing among alternative policies for mitigating global warming.

The scope of the external effects in the general model is admittedly unrealistic since global warming does not affect every consumer and firm worldwide. As we saw in the previous chapters, however, the pareto-optimal rules for the general model are easily modified if some people or firms are unaffected by the externality. Also, some forms of industrial pollution may approximate the general model within a small geographic region, such as water pollution by firms situated on a lake or bay.

The main advantage of the general model is that it is easily compared with our models of aggregate consumption and production externalities in Chapters 6 and 7. As it turns out, the policy rules are virtually identical in form.

Following the notation of Chapter 2, let

X_{hg} = consumption of good g by person h, where $g = 1,\ldots, G$ and $h = 1,\ldots, H$
V_{hf} = factor f supplied by consumer h (measured negatively) where $f = 1,\ldots, F$; $h = 1, \ldots, H$
r_{gf} = factor f used in the production of good g, $g = 1,\ldots, G$,[2] $f = 1,\ldots, F$
X^g = the aggregate output of good g, $g = 1,\ldots, G$.

Assume that the aggregate quantity of factor i (e.g., fuel) used in production enters the utility function of every consumer and every firm in the economy because its use contributes to global warming, P, that affects all agents. Let

$$P = P\left(\sum_{g=1}^{G} r_{gi}\right) = \text{global warming as a function of the aggregate use of factor i by the firms.} \tag{8.1}$$

$$X^g = \phi^{*g}\left(r_{gf}; r_{gi}; P\right) = \phi^g\left(r_{gf}; r_{gi}; \sum_{g=1}^{G} r_{gi}\right) \tag{8.2}$$
$$g = 1,\ldots,G; f = 1,\ldots,i-1, i+1,\ldots,F$$

$$U^h = U^{*h}\left(X_{hg}; V_{hf}; P\right) = U^h\left(X_{hg}; V_{hf}; \sum_{g=1}^{G} r_{gi}\right) \tag{8.3}$$
$$h = 1,\ldots,H; g = 1,\ldots,G; f = 1,\ldots,F$$

where $\phi^g(\,)$ = the production function for X^g, and $U^h(\,)$ = the utility function of person h. Notice that each production function, $\phi^g(\,)$, incorporates each firm's "personal use" of factor i, r_{gi} (e.g., using fuel to provide heat and to run engines, motors, and other machines), as well as the pollution externality $\sum_{g=1}^{G} r_{gi}$.

The usual assumptions about aggregate externalities apply:

$$\partial\phi^{*g}/\partial r_{ji} = \partial\phi^g/\partial\sum_{g=1}^{G} r_{gi} \quad j \neq g \tag{8.4}$$

2. The burning of gasoline by individuals' automobiles and of oil and natural gas to heat their homes can be thought of as a factor of production in the provision of automotive and housing services that individuals consume, two of the X's.

$$\partial\phi^{*g}/\partial r_{ji} = \partial\phi^{g}/\partial r_{ji} + \partial\phi^{g}/\partial \sum_{g=1}^{G} r_{gi} \quad j = g \tag{8.5}$$

$$\partial U^{h}/\partial r_{gi} = \partial U^{h}/\partial \sum_{g=1}^{G} r_{gi} \quad g = 1,\ldots,G \tag{8.6}$$

Society's problem is to maximize social welfare subject to the production constraints and market clearance:

$$\max_{\{X_{hg};V_{hf};X^{g};r_{gf};r_{gi}\}} w\left[U^{h}\left(X_{hg};V_{hf};\sum_{g=1}^{G} r_{gi}\right)\right]$$

$$X^{g} = \phi^{g}\left(r_{gf};r_{gi};\sum_{g=1}^{G} r_{gi}\right) \quad g = 1,\ldots,G$$

$$\text{s.t.} \quad \sum_{h=1}^{G} X_{hg} = X^{g} \quad g = 1,\ldots,G$$

$$\sum_{h=1}^{H} V_{hf} = \sum_{g=1}^{G} r_{gf} \quad f = 1,\ldots,F \text{ (including } i\text{)}$$

The model can be solved in the usual manner by defining Lagrangian multipliers for each of the production and market clearance constraints and taking derivatives of the resulting Lagrangian with respect to all the variables and the multipliers.

Generating and manipulating the first-order conditions so that they have the standard interpretations is tedious and will be left for the interested reader. We will simply note the principal results, which are entirely familiar thanks to the two dichotomies that apply to all first-best models. In particular, the usual interpersonal equity and pareto-optimal conditions are required to achieve a social welfare maximum at the bliss point.

The Interpersonal Equity Conditions

The government should redistribute one good or factor lump sum to equalize its social marginal utility of consumption (supply) across all consumers. The interpersonal equity conditions have the standard form:

$$\frac{\partial W}{\partial U^{h}}\frac{\partial U^{h}}{\partial V_{h1}} = \quad \text{all } h = 1,\ldots,H \tag{8.7}$$

with factor 1 chosen for redistribution. Assuming the interpersonal equity conditions hold, then the usual pareto-optimal conditions hold at the bliss point.

The Pareto-Optimal Conditions

The Purely Private Goods and Factors

The pareto-optimal conditions for all the purely private goods and factors have the standard form. For example,

$$\text{MRS}^{h} = \text{MRT} \quad h = 1,\ldots,H \tag{8.8}$$

for any two goods or factors.[3] They can be achieved by competitive markets without any government intervention.

The Externality

The pareto-optimal condition for the eternality-generating factor i also has the standard form expressed in terms of good 1. The difference between the private supply, $\text{MRS}^{h}_{i,1}$, and the private use, $\text{MRT}^{g}_{i,1}$, should equal the aggregate external effects of factor i on the margin. The only modification from previous models is that the external effects apply to all consumers and all firms. The pareto-optimal condition has the form

$$\text{MRS}^{h}_{i,1} - \text{MRT}^{g}_{i,1} = -\sum_{h=1}^{H}\text{MRS}^{h}_{P,1} + \sum_{g=1}^{G}\text{MRT}^{g}_{P,1} \tag{8.9}$$

Because the aggregate external effects on the right-hand side (RHS) are independent of the firm using factor i, a single Pigovian tax on the use of factor i can achieve condition (8.9), with

$$t_i = -\sum_{h=1}^{H}\text{MRS}^{h}_{P,1} + \sum_{g=1}^{G}\text{MRT}^{g}_{P,1} \tag{8.10}$$

and $P_1 = q_1 = 1$, the numeraire. In the context of global warming, the same tax on carbon emissions should be applied to every source of emissions worldwide (or the same tax on CO_2 emissions, if that is the gas targeted to be reduced). The tax is commonly referred to as a harmonized carbon tax.

The standard supply and demand analysis applies as well to factor i, as depicted in Fig. 8.1. Aggregate use of factor i, R_i, is on the horizontal axis. S represents the horizontal summation of each consumer's $\text{MRS}^{h}_{i,1}$ in supply.

Similarly, D^{priv} represents the horizontal summation of each firm's private use $\text{MRT}^{g}_{i,1}$. D^{soc} corrects D^{priv} by vertically subtracting the aggregate marginal external diseconomy at every aggregate R_i.

At the optimum, the difference between the private MRS and MRT, represented by the vertical distance $D^{\text{priv}} - S$, just equals the value of the marginal external effects

$$-\sum_{h=1}^{H}\text{MRS}^{h}_{P,1} + \sum_{g=1}^{G}\text{MRT}^{g}_{P,1}$$

represented by the vertical distance $D^{\text{priv}} - D^{\text{soc}}$. The Pigovian tax, t_i, drives the appropriate wedge between the producer demand price, P_i^{opt}, and the consumer supply price, q_i^{opt}, at the optimum. Thus, the only modification of our

3. Recall that MRT is interpreted as a marginal rate of transformation for two goods, a marginal product for a good and a factor, and a marginal rate of technical substitution for two factors.

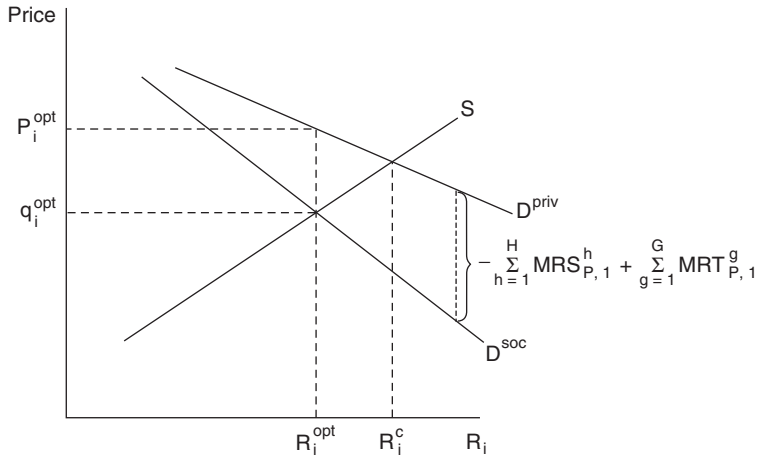

FIGURE 8.1

Chapters 6 and 7 analyses required by the consumption–production externality is in the terms reflecting the extent of the marginal damage, which now include both consumers and producers. Otherwise, the analysis is identical to that of the earlier models, not only for aggregate externalities but also for all other forms of externalities, as could easily be verified.

The only wrinkle regarding the harmonized tax is that fossil fuels are already taxed in many countries, such as taxes on gasoline used by automobiles. In these cases, the government should levy an additional tax as needed to raise the overall tax rate to the harmonized tax rate (or provide a subsidy to reduce the overall tax rate to the harmonized tax rate.)

A final point is that the same alternative solutions described in Chapter 7 are open to the government should it choose not to levy a direct tax on the source of the externality. For example, firms could be subsidized for reducing the amount of carbon emissions below the uncontrolled amount. In summary, the presence of a combined consumption–production externality changes none of the policy insights gained from analyzing the simpler consumption and production externalities.

LEGISLATING POLLUTION STANDARDS

Societies cannot be expected to achieve, or even approximate, the full social optimum when combating any significant pollution problem, much less one with all the potential third-party effects of global warming. The main problem comes in trying to evaluate the marginal benefits (MBs) of pollution reduction (the distance between D^{priv} and D^{soc} in Fig. 8.1). This is especially true for the portion of the MBs received by the consumers. In general, researchers and policy makers face three serious handicaps in determining the benefits.

The first is the enormous uncertainty about the harm caused by various pollutants, especially the conventional pollutants. Which pollutants are carcinogens, and at what concentrations? What is the precise relationship between the concentrations of the various pollutants in the atmosphere and the resulting increase in morbidity or mortality? Definite answers to questions such as these must await further scientific research.

Second, even if the effects of all pollutants were known, how does one evaluate the costs of pollution (benefits of less pollution). What values should be placed on such things as decreased visibility, increased morbidity, and loss of life? Economists have developed some ingenious survey and indirect market price techniques to estimate the benefits of reducing pollution. Despite their ingenuity, however, the benefit estimates from these techniques remain problematic and controversial. Also, the government must aggregate each person's marginal loss to arrive at the aggregate marginal damage on which the pollution tax is to be levied, and do so without the benefit of markets in which people are forced to reveal their preferences for cleaner air or water.[4]

A final problem, noted in Chapter 7, is that the government must measure the MBs at the optimum, not at the original preintervention equilibrium. In terms of Fig. 8.1, the tax should equal the divergence between D^{priv} and D^{soc} at R_i^{opt}, not at R_i^c. Even if the MBs at R_i^c were known with reasonable accuracy, their value at R_i^{opt} may well be subject to great uncertainty, especially if R_i^{opt} is far from R_i^c. (A trial-and-error process may discover R_i^{opt}, however, a point discussed in Chapter 7.)

All these problems are present with global warming. Carbon dioxide and the other GHGs remain in the atmosphere for hundreds of years, thereby requiring estimates of damages caused by global warming into the distant future.

4. Cropper and Oates (1992) have an excellent discussion of attempts to measure the benefits (and costs) of pollution.

These estimates are especially problematic because climate scientists cannot be sure how much temperatures will rise for any given sustained emissions rate of GHGs, or how much melting of the polar ice caps will occur given the estimated rise in temperature, and therefore how much the sea level will rise. Projections are all over the place. Whatever projection one settles on, the various marginal benefits and costs across consumers and firms are so extensive that attaching reasonably accurate numbers to the sum of the MRS's and MRT's is problematic in the extreme. Then adjusting these estimates to their sum at the optimum adds another huge element of guesswork. We will return below to the special problems of estimating costs and benefits far into the future.

Given all these difficulties, governments have thrown in the towel and turned to a second-best "standards" approach for controlling pollution. First they somewhat arbitrarily select a desired target level for each pollutant. Then the economic goal becomes meeting the legislated target at the lowest possible cost. The international attempts through the United Nations to reduce global warming are another instance of this approach.

The Kyoto Protocol and the Copenhagen Accord

The United Nations began efforts to forge an international agreement to reduce GHG emissions in 1992 under the auspices of its Framework Convention on Climate Change (UNFCCC). The goal was to stabilize the concentration of the GHGs in the atmosphere and thereby halt the increase in global warming. Its efforts reached fruition in December of 1997 with the signing of the Kyoto Protocol. Thirty-nine developed countries, those of the Organization for Economic Cooperation and Development, Eastern Europe, and the former Soviet Union agreed to reduce their GHG emissions to achieve an overall reduction of global emissions 5% below global emission levels of 1990 during the period from 2008 to 2012. The countries were assigned individual reduction targets ranging from 8% (23 countries) reductions to a 110% increase (Iceland). The US target was a 7% reduction. The countries took these pledges back to their own governments for approval, and the Protocol was ratified in February, 2005.

In addition to the targeted reductions, the Protocol recommended that the reductions be achieved by means of marketable emissions permits, with each permit allowing one ton of carbon emissions. Permits would be distributed based on 1990 emissions and then traded among countries. Any country's emissions each year could not be more than the number of permits that it owned. The marketable permits were viewed as the least cost means of reaching the overall targeted reduction. (We will analyze marketable permits in the next section.)

The Kyoto Protocol was essentially doomed by the time it was ratified, for two main reasons. One is that it did not apply to the developing countries, particularly the large developing countries such as China and India. The other is that the United States dropped out of the Protocol in 2001, and US emissions of CO_2 represented 32% of total CO_2 emissions. By 2005, the countries still covered by the Protocol accounted for only 30% of global emissions. Moreover, only the European Union among the 39 countries adopted the suggested use of marketable permits to reduce CO_2 emissions, a program called the European Trading Scheme (ETS). The ETS covered only 8% of global CO_2 emissions, however. As it happened, global GHG emissions had fallen by 2008−2012, but this was almost entirely due to reduced production as a result of the worldwide Great Recession that hit at the end of 2007. The Protocol was essentially ignored, other than by the EU.[5]

With the 2012 deadline fast approaching, 120 nations met in Copenhagen in December of 2009 to generate a new agreement on GHG emissions to apply post-Kyoto. The results were not encouraging. Many nations left without agreeing to anything. In order to salvage something from the meeting, the United States and a few other countries drew up an accord that set a target of holding the increase in global warming to 2 °C (3.6 °F). The developed countries that signed the original Kyoto Protocol in 1997 were to submit targeted reductions in CO_2 emissions by January 31, 2010. The Copenhagen Accord also included some developing countries, notably China, which were to begin implementing what were termed "mitigating strategies" and submit their plans for these strategies, along with their targeted CO_2 emissions reductions, also by January 31, 2010. There was no suggested method in the Accord for meeting the targeted reductions at least cost. The Accord also established a Copenhagen Green Climate Fund by which the developed countries would provide subsidies to the developing countries to help the latter reach their targets without undue costs.

Most countries did submit targeted reduction plans, but they amounted to little more than empty promises because there was no enforcement mechanism put in place. For example, the Obama administration committed the United States to a 17% reduction of its GHG emissions below 2005 levels by 2020, followed by a 45% reduction from 2005 levels by 2030, and an 85% reduction below 2005 levels by 2050. While this sounds impressive, the administration had no chance at all of getting Congress to accept this commitment. In truth, there is no international agreement in place to reduce

5. Details of the Kyoto Protocol are available on the UN web site: http://unfccc.int/kyoto_protocol/items/2830.php. Two additional sources for the difficulties Kyoto encountered are Buchner (2002) and Nordhaus (2005).

GHG emissions, nor is one on the horizon as this is written in 2014. At the same time, climate scientists believe that there is some probability that unchecked GHG emissions could raise average temperatures by 6.7 °C (12 °F) by 2100, an increase that would lead to catastrophic loss of life.[6]

The Kyoto Protocol target of a 5% reduction of GHG emissions below 1990 levels by 2008–2012 and the Copenhagen Accord target of holding the increase to global warming to 2.6 °C by 2100 are examples of the standards approach to reducing pollution. Neither target was arrived at as a result of an attempt to estimate the marginal benefits and costs of reducing GHG emissions and thereby establish an optimal strategy for reducing emissions. Rather, they were seen as reasonable targets to strive for that would likely be beneficial on net. The economic issue, then, is how to meet these legislated standards at the lowest possible cost.

Cost Minimizing under the Standards Approach to Reducing Pollution

To address the least cost issue, we have to modify our first-best model to incorporate the standards approach. Fortunately, the modification turns out to be straightforward. Minimizing the opportunity costs of achieving a given pollution standard is formally equivalent to maximizing social welfare subject to the additional constraint that the standard is satisfied. The reason is that the opportunity costs are simply the losses in social welfare from satisfying the constraint.

Adding the pollution constraint makes the analysis second best so long as the target level of pollution differs from its first-best optimum level. Nonetheless, we will briefly sketch out the constrained model here because of its relevance to pollution policy and because the solution to the constrained social welfare optimum problem is also a single tax, the properties of which are virtually identical to the unconstrained Pigovian tax.

To analyze the constrained social welfare optimum, let us continue with the same aggregate consumption–production externality as above, in which

$$\text{Global warming } (P) = P\left(\sum_{g=1}^{G} r_{gi}\right) \tag{8.11}$$

Assume the government arbitrarily targets the pollution standard at $\overline{P}$. Since P is assumed to be a monotonic function of $\sum_{g=1}^{G} r_{gi}$, reinterpret the constraint to be

$$\overline{R}_i = \sum_{g=1}^{G} r_{gi} \tag{8.12}$$

where $\overline{R}_i$ corresponds to $\overline{P}$. Assume further that the pollution target is the only additional constraint in an otherwise first-best policy environment. In other words, the formal aggregate externality model above applies with the addition of the pollution constraint.

The Lagrangian of the formal problem becomes

$$\begin{aligned}\max_{\{X_{hg};\, V_{hf};\, X^g;\, r_{gf};\, r_{gi}\}} L = {} & W\left[U^h\left(X_{hg}; V_{hf}; \sum_{g=1}^{G} r_{gi}\right)\right] \\ & + \sum_{g=1}^{G} \mu_g \left(X^g = \varphi^g\left(r_{gf}; r_{gi}; \sum_{g=1}^{G} r_{gi}\right)\right) \\ & + \sum_{g=1}^{G} \delta_g \left(\sum_{h=1}^{H} X_{hg} - X^g\right) \\ & + \sum_{f=1}^{F} \pi_f \left(\sum_{h=1}^{H} V_{hf} - \sum_{g=1}^{G} r_{gf}\right) + \lambda\left(\overline{R}_i - \sum_{g=1}^{G} r_{gi}\right)\end{aligned}$$

By inspection, the first-order conditions for all pure private goods and factors are identical to those of the unconstrained model. As before, the government need not intervene in any market except the market for factor i. Similarly, the interpersonal equity conditions remain unchanged since they are unaffected by r_{gi}. The only difference is the first-order condition for the r_{gi}, which now includes the term $-\lambda$, the multiplier applied to the pollution constraint. λ equals the marginal increase in social welfare from relaxing the pollution constraint, measured at the second-best optimum. Alternatively, λ is the marginal social cost of reducing the use of R_i to $\overline{R}$.

The pareto-optimal condition for r_i, expressed in terms of factor 1, is now

$$\text{MRS}^h_{i,1} - \text{MRS}^g_{i,1} = -\sum_{h=1}^{H} \text{MRS}^h_{P,1} + \sum_{g=1}^{G} \text{MRT}^g_{P,1} + \lambda/\pi_1 \tag{8.13}$$

which differs from Eqn (8.9) only by the presence of the term λ/π_1. The terms on the left-hand side (LHS) represent the private marginal rates of substitution in use and supply. The first two terms on the RHS represent the marginal social costs of the pollution externality to the consumers and producers. The final term on the RHS is an additional marginal social cost from imposing the resource constraint on the solution. (The division by π_1 expresses the loss of social welfare in terms of good 1, since $\pi_1 = \frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial V_{h1}}, h = 1, \ldots, H,$ from the first-order condition for V_{h1}.)

6. One particular difficulty in reaching any international agreement on reducing global warming is the principle established by the Treaty of Westphalia in 1648, that no country can be forced to accept an international agreement if it does not choose to do so. Details of the Copenhagen Accord are available on the UN's Web site: unfccc.int/resource/docs/2009/cop15/eng/l07.pdf. The country's pledged CO_2 reductions are on http://unfccc.int/home/items/5264.php (United States and European Union) and http://unfccc.int/home/items/5265.php (China).

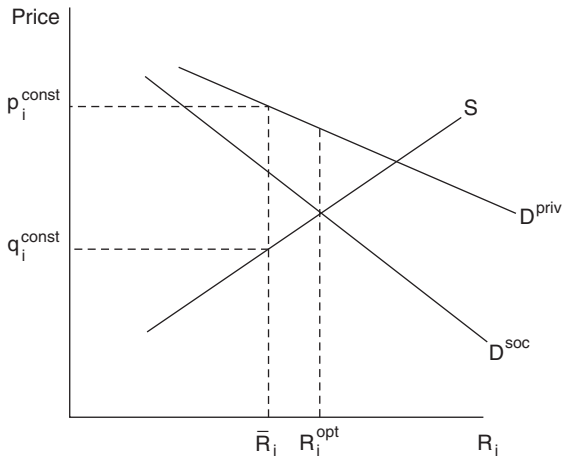

FIGURE 8.2

Given the legislated constraint $\overline{R}(\overline{P})$, therefore, the divergence between the private use MRT and MRS now equals the aggregate external effects on the margin plus the social marginal cost (MC) of the constraint at $\overline{R}_i$. Note also that the RHS is independent of g; it does not matter which firm pollutes on the margin. Therefore, a single Pigovian-style tax can achieve the second-best optimum, as illustrated in Fig. 8.2.

R_i^{opt} is the unconstrained first-best optimum and $\overline{R}_i$ is the pollution standard imposed by the government. The producer and consumer prices as the constrained optimum are p_i^{const} and q_i^{const}, with the difference between them equal to the tax t_i. The government imposes $\overline{R}_i$ by adjusting the tax until $R_i = \overline{R}_i$. As in Fig. 8.1, the vertical distance between D^{priv} and D^{soc} equals the third and fourth terms in Eqn (8.13), the direct marginal social costs on the producers and consumers caused by the externality. The additional distance between D^{soc} and S at $\overline{R}_i$ is a measure of λ/π_1, the marginal social cost of the pollution constraint. Notice that λ/π_1 is a residual cost, in effect. Once the government sets the tax to obtain $\overline{R}_i$, the tax automatically represents the full marginal social cost of using factor i at the second-best optimum. The diagram also indicates that the marginal constraint cost is zero only if $\overline{R}_i = R_i^{opt}$.[7]

The final point is that this tax must minimize the opportunity costs of achieving the resource constraint. This could be demonstrated by setting up a cost minimization problem but that is not necessary. Since the tax satisfies the pareto-optimal conditions of the constrained social welfare maximization problem, it must be cost minimizing in a production sense. If production were inefficient, a reallocation of resources could increase outputs without additional resources and the bonus could be given to consumers to increase social welfare. But this contradicts the fact that social welfare is maximized given the arbitrary pollution standard. $\overline{R}_i(\overline{P})$ may be a terrible choice, far from the first-best optimum, but given that society chose it, taxing the use of resource i such that $\overline{R}_i$ is met is the least cost way of achieving $\overline{R}_i(\overline{P})$.

Taxing to meet a pollution standard may be a least cost strategy, but it is not necessarily the one chosen by governments. The United States has long favored a regulatory approach to reducing pollution, called a command and control (CAC) approach, under which each polluter is required to reduce pollution by a given amount or to install certain devices or use certain technologies to reduce pollution. A familiar example is the pollution control equipment required on all automobiles sold in the United States. As noted earlier, the Kyoto Protocol recommended a system of marketable permits to control the aggregate amount of GHG emissions. A system of marketable permits is often referred to as a cap-and-trade approach: The aggregate emissions are capped by the total amount of permits issued and the permits, once issued, can be traded in an organized permits market from sources whose emissions are less than their number of permits to sources whose emissions exceed their number of permits. We want to compare these two strategies with the cost-minimizing tax strategy for meeting a given pollution standard, beginning with the CAC approach.

The CAC Approach

The CAC regulatory approach to industrial polluters that dominates US antipollution policy is seriously flawed relative to a tax policy. In principle, regulation can be designed to be equivalent to the tax for any given standard, but only if it duplicates the exact pattern of resource use and production occasioned by the tax. This is clearly not practicable. Instead, regulation invariably takes the form of simple rules that are decidedly worse than the tax. For example, the Kyoto Protocol called for a 5% reduction in GHG emissions. The cost-minimizing solution is to tax all emitters of GHG emissions until the emissions are reduced by 5% in the aggregate. If direct regulation is used instead, there may be little choice other than to dictate a 5% reduction in GHG emissions by each emitter. The government may even dictate the methods used to achieve the reduction, as the US federal government typically does in its policies for reducing air and water pollutants. The regulatory approach may achieve the desired 5% reduction, but it does so at opportunity costs far in excess of the tax policy.

The flaw in the regulatory strategy is that it ignores the differences in costs and substitution possibilities across firms. The tax strategy, in contrast, exploits these differences in a least cost manner. Asking firms to reduce

7. As in the unconstrained case, whether a single tax or separate taxes on each source of the pollution is least cost depends on whether the externality is aggregate (single tax) or individualized (separate taxes).

their pollution equally in the name of fair play may strike some people as equitable. If so, it is a very costly notion of equity and one that has no standing in the quest to maximize social welfare. Yet, this notion of even-handed fair play appears to dominate antipollution policy in the United States.

Further intuition for the cost advantage of a pollution tax over the CAC approach can be gained from the simple textbook least cost production rule. Consider CO_2 emissions from the burning of fossil fuels.

Think of some firm producing its output (Q) with the use of capital (K), labor (L), and fuel (F), according to the production function:

$$Q = Q(K, L, F)$$

The least cost production rule says that to produce any given amount of Q at least cost, the firm must equalize the ratios of marginal product to price across all three inputs:

The ratios represent the extra output per dollar from using each of the factors. If the ratios are unequal, the firm should substitute toward the factors with the higher ratios (the higher marginal output per dollar) until the ratios are equalized.

The least cost production rule highlights the general principle that a quantity complaint ("there is too much CO_2 emission") is symptomatic of a pricing problem. As such, a tax (or equivalent pricing mechanism) gets right to the heart of the problem. A tax on fuel forces an additional price on polluters that reflects the marginal damage of burning fossil fuels. If the firms have driven ratios of the marginal products of K, L, and F to their prices before the tax, the value of the ratios with the tax is

$$\frac{\mathrm{MP}_K}{P_K} = \frac{\mathrm{MP}_L}{P_L} > \frac{\mathrm{MP}_F}{P_F}$$

Firms now have an incentive to substitute away from fuel and toward capital and labor to minimize their production costs. In other words, the tax combats pollution by appealing to the same profit motive that led to the pollution in the first place. It also gives firms flexibility to respond to the tax depending on their ability to substitute away from fuel. For example, firms that are highly profitable and have difficulty substituting for fuel may simply pay the tax and continue to emit CO_2 as before. Other firms that can easily find substitutes for the fuel they are using, such as switching from fossil fuel to solar energy, may find that substituting capital and labor to provide the solar energy is less expensive than paying the tax and polluting.

The cost-minimizing (social welfare optimizing) properties of taxing pollution at the source to meet a pollution standard should now be clear. A correctly designed tax forces firms to consider the full social costs of their decisions, and the tax is easy to design correctly. Simply adjust the tax until the standard is met. In addition, a tax permits each firm to respond as flexibly as possible to these social costs. The tax raises the costs for each polluting firm because they now pay more to use a resource that they previously used as part of their cost minimizing strategy. Each firm reacts by trying to minimize these additional costs. With each firm seeking its own least cost reaction to the tax, society's costs of reaching the standard are minimized in the aggregate. This is the essence of a tax or pricing strategy for combating pollution.[8]

A second practical drawback to regulation, stressed long ago by Mills (1967), is that the incentive structure is literally backward. With a harmonized carbon tax on fossil fuels, firms have an incentive to substitute away from fossil fuels to minimize their tax burdens. With regulation, firms have a profit-motivated incentive to cheat because the direct price of burning fossil fuel has not changed. Rational firms will weigh the cost advantages of continuing to pollute against the probability of being caught and the penalty for cheating. Moreover, it is up to the government prosecutors to bring suit, and the burden of proof is on the prosecution. A prosecutorial approach is highly problematic in an international setting. With taxes, in contrast, the firms bear the burden of proving that they deserve a lower tax bill because they have reduced their carbon emissions.

Direct regulation makes sense as a standby weapon for short-term emergencies. If air pollution becomes extremely dangerous because of unusual atmospheric conditions, then a temporary ban on some air pollutants may be the only effective short-term solution. Also, if the United States had maintained its original goal of zero water pollution, then the choice of taxes versus regulation is irrelevant. The only way to achieve zero pollution in the aggregate is for each polluter to stop polluting entirely—taxes or CAC necessarily lead to the same solution firm by firm. But this situation is not relevant in the context of global warming.

Marketable Permits

The recommendation in the Kyoto Protocol for a system of marketable permits to control GHG emissions was not a new idea at the time. Governments had used a system of marketable permits to control pollution within their own countries. A notable example was the United States adoption of marketable permits in 1990 to reduce the emissions of sulfur dioxide by the electric utilities, the first time the United States used a pricing strategy instead of the CAC approach to reduce any pollutant.

Marketable permits are equivalent to pollution taxes in principle. To see why, recall the model from Chapter 7 in

8. Production possibilities must also remain convex in the presence of the externality. Refer to the discussion of this point at the end of Chapter 7.

which a firm (firm 1) produces a by-product z_1 that affects the production of all other firms. Part of the discussion there compared various kinds of taxes and subsidies for reducing the firms' output of z_1 by seeing how each affects the profit function of firm 1.

Under a straight tax on z_1, t_z, the profit function of the firm is

$$\text{Profits} = \sum_{n=1}^{N} P_n X_{\text{ln}} - t_z z_1 \tag{8.14}$$

where X_{ln} are the purely private goods and factors supplied and purchased by firm 1. Think of z_1 as tons of CO_2 emissions for our current purposes.

Under a marketable permit scheme, a government or international agency determines the total amount of permits to be issued, with each permit allowing one ton of CO_2. Therefore, the number of permits equals the total amount of emissions allowed. The permits can then be auctioned off to the emitting producers or distributed to them free of charge as determined by some distribution formula. The producers must buy permits for whatever amount of CO_2 they choose to emit. Define z_p as the number of permits purchased by a firm. The permits, once distributed by auction or by formula, are traded in a national or international market that has established a price of P_p for each permit. Consider first auctioning off the permits.

Auctioning the permits—Let firm 1 represent a CO_2 emitter. The firm's profit function under the marketable permit scheme becomes

$$\text{Profits} = \sum_{n=1}^{N} P_n X_{\text{ln}} - P_P z_p$$
$$\text{s.t. } z_1 \leq z_p$$

Assuming the constraint is binding (firms will not buy permits they do not intend to use), the profit function is

$$\text{Profits} = \sum_{n=1}^{N} P_n X_{\text{ln}} - P_P z_1 \tag{8.15}$$

Therefore, the permits and tax are identical from the perspective of firm 1, providing $t_z = P_p$.

That the tax and permit price are equal for a given reduction in pollution can be seen from Fig. 8.2. Think of $\overline{R}_i$ as the legislated target for CO_2 emissions, and $t_i = t_z$. The tax is set such that it induces firms to emit exactly the target level of emissions, $\overline{R}_i$. Thus, if the government issues a total number of permits equal to $\overline{R}_i$ and the permits are traded in a competitive market, the equilibrium price P_p must equal the tax. If $P_p > t_z$, the firms would want to purchase fewer than $\overline{R}_i$ permits, the permit market would be in excess supply, and the price would fall to t_z. Similarly, P_p would be driven up to t_z if it were originally less than t_z and the permit market were in excess demand. With $P_p = t_z$, the permit and tax schemes are identical.

Distributing the initial permits free of charge—Suppose each firm, including firm 1, receives $\overline{P}$ permits initially. Firm 1's profit function is now

$$\text{Profits} = \sum_{n=1}^{N} P_n X_{\text{ln}} - P_P\left(z_1 - \overline{P}\right) \tag{8.16}$$

or

$$\text{Profits} = \sum_{n=1}^{N} P_n X_{\text{ln}} - P_P z_1 + P_p \overline{P} \tag{8.17}$$

Distributing the permits initially free of charge is kinder to the producers, but it does not change the first-order conditions for profit maximization. The permits are still equivalent to the tax on the margin. Also, the subsidy to the firms is different from the subsidy solution described in Chapter 7, since any new entrant who will emit CO_2 has to purchase permits from existing firms. It does not receive a subsidy to enter the market unless additional permits are granted each year, in which case it may receive a subsidy. But this would defeat the government's purpose of holding constant the emission of CO_2.

Practical considerations may favor permits or taxes in certain applications, however. Permits have the advantage of assuring that the legislated target is met at the outset of the program, whereas taxes have to be adjusted until the target is reached. Also, the price of permits automatically adjusts to the general level of inflation. A pollution tax, in contrast, would have to be increased every year to maintain the appropriate relative value of the tax. Countering these advantages is the possibility that existing firms could hoard the permits. We assumed above that $z_1 = z_p$. A firm could, however, buy more permits than it intended to use for the sole purpose of preventing other firms from obtaining them. Hoarding could be a very effective barrier to entry, either by preventing new firms from entering the market or by forcing existing firms to reduce their production or increase their costs if they cannot obtain their profit-maximizing number of permits.

Marketable Permits, Taxes, and Uncertainty

Another consideration that could favor marketable permits over taxes relates to the uncertainties surrounding the MBs and MCs of reducing almost any kind of pollution. The optimal amount of pollution occurs when the marginal benefits and costs are equal, as depicted in Fig. 7.3. The government is unlikely to be able to measure either the MBs or the MCs for any pollutant with much precision, however. At best it might only have some intuitive sense of the relative shape of the MB and MC curves. For example, the government might reasonably assume that the MB curve becomes quite steep for many toxic substances at

some threshold level of pollution reduction simply because high concentrations of these substances can be so dangerous. (The MBs of reducing pollution are the same as the marginal damages of increasing pollution.) Alternatively, the government might just as reasonably assume that the MC becomes quite steep for all pollutants beyond some very high level of pollution reduction. Reducing pollution any more beyond that point may require extremely costly abatement techniques. Even rough guesses about the relative shape of the MB and MC curves turn out to be important information for they can lead to a preference for marketable permits or taxes.

Suppose the government assumes that the MB curve is quite steep and the MC curve is quite flat in some region of pollution reduction, as is likely for toxic substances. Marketable permits are the preferred strategy under this scenario because it is better able to control the quantity of the pollutant, assuming it can be enforced. Pollution can be set at a safe level below the danger threshold. If this happens to be beyond the optimal MB = MC point, at least it avoids the steep portion of the MB curve without incurring large increases in cost. Using taxes, in contrast, runs the risk that the tax will be set too low, there will be too little pollution reduction, and society will remain exposed to dangerous concentrations of the pollutant.

Suppose, instead, that the government assumes the MC curve is quite steep and the MB curve is relatively flat in the region near the target level of pollution. This combination is highly likely with conventional pollutants if the target levels are fairly stringent. A pollution tax is the preferred strategy under this scenario since the tax allows tighter control over the costs of reducing pollution. Errors in setting the tax too high or too low have rather modest effects on marginal damages. Marketable permits, in contrast, run the risk of setting a much too stringent target, which sharply raises the MCs of pollution reduction without much offset from the MBs.[9]

The Preference for Taxes over Marketable Permits for CO_2 Emissions

Global warming brings up a number of other considerations that have led most economists to conclude that the Kyoto Protocol erred in recommending marketable permits to reduce GHG emissions. A harmonized carbon tax is likely to be a much better strategy, for a number of reasons. The first relates to the analysis in the preceding section. Both the MBs and MCs of reducing global warming are full of uncertainties. Nonetheless, we can be pretty sure that the MB

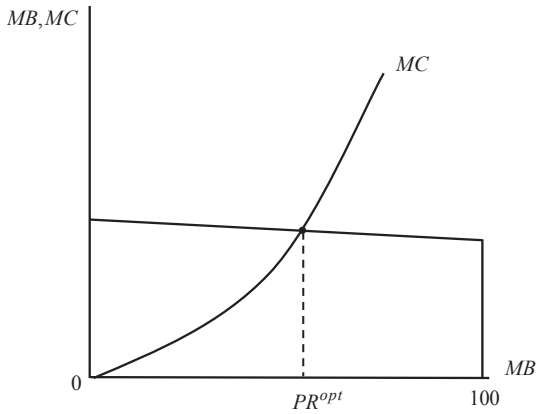

FIGURE 8.3

curve in any one year is fairly flat over virtually its entire range, as pictured in Fig. 8.3. The GHG emissions are a stock externality, because the amount of global warming depends on the total accumulation of the gases in the atmosphere and the gases remain in the atmosphere for a very long period of time. Therefore, whether emissions in any one year are reduced substantially or hardly at all makes little difference to the accumulated stock of gases in the atmosphere. Hence the MB curve has to be quite flat. In contrast, the MC curve is likely to have the usual fairly steep slope, since reducing GHG emissions in any context is quite costly for firms, and the costs increase rapidly with greater emission reduction. The combination of a flat MB curve and a relatively steep MC curve argues in favor of the harmonized tax to keep tighter control over the costs of the emissions reductions.

A second point is that nations are already quite familiar with taxes on carbon emissions since they are commonly used. An obvious example is the excise tax on gasoline, which is levied throughout the world. A harmonized carbon tax would have to take into account the taxes on carbon that already exist, since existing taxes would have to be adjusted up or down to be equivalent to a worldwide harmonized tax on each ton of carbon (or CO_2) emissions. But this is an easier exercise for most countries than establishing a permits, cap-and-trade market from scratch.

A third point is that the cap-and-trade systems that do exist have been vulnerable to wildly varying price swings for permits, which makes it difficult for firms to plan future investment strategies. For example, prices for permits under the European Union's ETS have varied from a high of $104 to a low of $23 per ton of carbon; prices under the United States' cap-and-trade system to control emissions of sulfur dioxide from electric utility plants were even more variable while the system was still operating as a national market, from a high of $1500 per ton of SO_2 in 2005 to a low

9. The general preference for quantity or price controls under uncertainty was first analyzed by Weitzman (1974).

of $70 in 1996.[10] It turns out that determining the proper amount of permits each year is a difficult judgment call given technical change and variations in macroeconomic conditions. This would certainly be true on a global scale. As noted above, a harmonized tax would have to be adjusted periodically as well, but the wild swings under permit prices can be avoided.

Finally, political considerations argue for a harmonized tax over marketable permits if monitoring firms' emissions within a country is difficult for an international agency to do, as would almost certainly be true in the case of CO_2 emissions. The problem with marketable permits is that they would first be distributed to each country, and then the country would decide how to allocate them to its various CO_2-emitting producers. Under this system, and assuming imperfect monitoring, the firms and the government have the same interests. The firm wants to claim that it has reduced its emissions so that it can be a seller rather than a buyer of permits. The country's government has the same interests, to be a seller of permits on the international market because it is an easy source of revenue. Therefore, the government's incentive is to take the firms at their word and not monitor them closely. They certainly have no interest in allowing an international agency to closely monitor their firms. Under a harmonized carbon tax, in contrast, a country's government and its emitting producers have opposite interests: the firms want to claim they have reduced emissions to save on taxes but the government wants the tax revenue from the firms. Hence the government has an incentive to closely monitor its own firms. This difference in incentives is probably enough by itself to prefer a harmonized tax over a cap-and-trade program since the monitoring of emissions at the firm level by an international agency is highly unlikely to be very effective, or indeed even acceptable to many nations (China refused to submit to monitoring under the Copenhagen Accord, even though it pledged a considerable reduction in GHG emissions).

DEFENSIVE ANTIPOLLUTION STRATEGIES

Pollution taxes, marketable permits, and CAC are strategies designed to reduce pollution at its source. In most cases involving pollution, those harmed by the pollution can also defend themselves to some extent. Pollution can also be removed after the fact; municipal treatment plants of polluted water and the United States' Superfund are examples.

Defending against the global warming caused by GHG emissions is also possible. One possibility is to replant the Brazilian rain forest and other forestation programs; trees and plants remove CO_2 from the atmosphere in the process of photosynthesis. Another option is for people to move away from coastal areas or to build seawalls to protect coastal cities and towns as the ocean level rises. Still another possibility is atmospheric engineering. Some scientists believe it would be possible to seed the atmosphere with sulfur-based compounds that would break down the accumulated GHG gases and reverse the global warming. This is highly controversial, however; other scientists believe it is little more than a pipe dream given current scientific knowledge. No doubt other possibilities exist as well to defend against global warming.

Removing pollution after the fact and defending oneself from pollution are certainly policies worth considering. Suppose someone invented a method for removing vast amounts of pollutants or hazardous wastes for only a few pennies. Clearly an antipollution policy would want to make liberal, perhaps exclusive, use of this technology. Recall the factory-laundry example from Chapter 7, which suggested that simply moving the laundry upwind from the factory may be less expensive and more effective than a Pigovian tax designed to reduce the factory's smoke emissions. Generally speaking, any technology or solution is attractive so long as it is "cheap enough." But whether the government should subsidize defensive antipollution strategies is a complex question that depends on the exact form of the externality, available policy options, and the nature of the defensive technologies.

All our models of pollution externalities so far in Chapters 6 through 8 have reached the same conclusion, that the government must reduce pollution at its source to achieve pareto optimality. These models, however, have not included the possibility of defensive measures by individuals or firms to reduce their exposure to pollution. A number of interesting questions arise once defensive measures are admitted:

1. Does the government still have to reduce pollution at the source or can it rely exclusively on defensive measures?[11]
2. Is a mixed strategy of reducing pollution at the source and removing pollution after the fact pareto optimal, in general?
3. Does the government have to subsidize defensive measures to achieve pareto optimality or can individuals and firms be counted on to respond optimally on their own?

We want to consider these questions in a general framework that could apply to many different kinds of pollution rather than just to global warming, although we will return to global warming at the end of the section.

10. Nordhaus (2005). The U.S. SO_2 prices are reported on p. 15 and the EU-ETS carbon prices on p. 8.

11. Superfund and waste-treatment efforts are obviously necessary to remove existing pollutants that resulted because no attempt was made to reduce them at their sources and that would remain harmful if not treated or removed. Our question is meant to be forward looking, when all strategies are possible.

To explore defensive strategies in a relatively simple framework, let us return to the single-source production externality model of Chapter 7, in which firm 1 produces a substance, z_1, that enters into all other firms' production functions. Think of z_1 as a pollutant that harms the firms. There are J firms and N goods and factors besides z_1, and all the other goods and factors, X_{jn}, are purely private.

To incorporate the possibility of defensive measures, assume that each of the other firms can reduce its exposure to z_1 with the use of factor k, according to the function $h^j(X_{jk})$. Firm 1 can also use factor k to reduce its own exposure to z_1 should it want to.

Firm j's exposure to the externality is

$$Z_{JI} = z_1 - h^j(X_{jk}) \quad j = 1, \ldots, J \tag{8.18}$$

with $\partial h^j/\partial X_{jk} > 0$.

Define the production functions for each of the J firms to be

$$f^j(X_{jn}; {}_{Z1} - h^j(X_{jk})) = 0 \quad j = 1, \ldots, J; n = 1, \ldots, N \tag{8.19}$$

This formulation assumes that input X_{jk} may have some direct use for firm j besides its use in reducing z_{j1}. An example might be an all-purpose cleaning agent or a cooling technology that has many uses to firms besides reducing the harmful effects of z_1.

Assume a one-consumer equivalent economy because the interpersonal equity conditions hold. Let $U(\sum_{j=1}^{J} X_{jn}) =$ the utility function of the single consumer, incorporating the N market clearance equations, with $\partial U/\partial X_{jn} = \partial U/\partial X_n$, all $J = 1,\ldots, N$. The consumer is indifferent about the identity of firm j. Note, also, that the consumer is unconcerned about the production of z_1 since it generates only a production externality.

The government's problem is

$$\max_{(x_{jn}; z_1)} U\left(\sum_{j=1}^{J} X_{jn}\right)$$
$$\text{s.t.} f^j(X_{jn}; z_1 - h^j(X_{jk})) = 0$$

with the corresponding Lagrangian:

$$\max_{(x_{jn}; z_1)} L = U\left(\sum_{j=1}^{J} X_{jn}\right) + \sum_{j=1}^{J} \lambda_j f^j(X_{jn}; z_1 - h^j(X_{jk}))$$

The first-order conditions are

$$\partial L/\partial X_{jm} = \partial U/\partial X_m + \lambda_j \partial f^j/\partial X_{jm} = 0 \quad j = 1, \ldots, J; m \neq k \tag{8.20}$$

$$\partial L/\partial X_{jk} = \partial U/\partial X_k + \lambda_j(\partial f^j/\partial X_{jk} - \partial f^j/\partial X_{jl} \cdot \partial h^j/\partial X_{jk}) = 0 \quad j = 1, \ldots, J \tag{8.21}$$

$$\partial L/\partial z_l = \lambda_l \partial f^l/\partial z_l + \sum_{j=2}^{J} \lambda_j \partial f^j/\partial z_{jl} = 0 \tag{8.22}$$

Express the first-order conditions as ratios in terms of X_1, and assume X_1 is a good for purposes of interpretation. From conditions (8.20) for m and 1,

$$U_m/U_l = \frac{\frac{\partial f^j}{\partial X_{jm}}}{\frac{\partial f^j}{\partial X_{jl}}} \quad j = 1, \ldots, J; m \neq k \tag{8.23}$$

Adding the possibility of defensive measures does not change the result that the government need not intervene in the markets of the purely private goods and factors that are unrelated to the externality. They can be marketed competitively, with the consumer and producer prices equal for all the private goods: $q_m = p_m$, $m \neq k$.

Consider, next, the allocation of z_1 expressed in terms of good 1. Dividing Eqn (8.22) by the first-order conditions for good 1 in Eqn (8.20) and selectively eliminating all the λ_j in the usual manner yields

$$\frac{\frac{\partial f^1}{\partial z_l}}{\frac{\partial f^l}{\partial X_{ll}}} + \sum_{j=2}^{J} \frac{\frac{\partial f^j}{\partial z_{jl}}}{\frac{\partial f^j}{\partial X_{jl}}} = 0 \tag{8.24}$$

Condition (8.24) can be satisfied by a pollution tax on firm 1, t_z, such that

$$t_z/q_l = -\sum_{j=2}^{J} \frac{\frac{\partial f^j}{\partial z_{jl}}}{\frac{\partial f^j}{\partial X_{jl}}} \tag{8.25}$$

where t_z is the standard Pigovian tax, equal to the aggregate damage of z_1 on the margin to all the affected firms. Faced with the optimal tax, firm 1 sets the tax equal to the value of marginal product of z_1 in its production of X_{11}:

$$t_z = \left(\frac{\frac{\partial f^l}{\partial z_l}}{\frac{\partial f^l}{\partial X_{ll}}}\right) q_1, \quad \text{or} \quad t_z/q_l = \frac{\frac{\partial f^l}{\partial z_l}}{\frac{\partial f^l}{\partial X_{ll}}}$$

Therefore,

$$t_z/q_l = \frac{\frac{\partial f^l}{\partial z_l}}{\frac{\partial f^l}{\partial X_{ll}}} = -\sum_{j=2}^{J} \frac{\frac{\partial f^j}{\partial z_{jl}}}{\frac{\partial f^j}{\partial X_{jl}}} \tag{8.26}$$

as required for pareto optimality. The answer to the first question above, then, is that the government should continue to tax the pollutant at its source in the presence of defensive measures.

Consider, next, the optimal allocation of the defensive factor X_k expressed in terms of good 1. Dividing Eqn (8.21)

by the first-order conditions for good 1 in Eqn (8.20) and selectively eliminating all the λ_j in the usual manner yields

$$U_k/U_l = \frac{\frac{\partial f^j}{\partial X_{jk}}}{\frac{\partial f^j}{\partial X_{jl}}} - \left(\frac{\frac{\partial f^j}{\partial z_{jl}}}{\frac{\partial f^j}{\partial X_{jl}}}\right)\frac{\partial z_{jl}}{\partial X_{jk}} \quad j = 1, \ldots, J \tag{8.27}$$

The RHS of Eqn (8.27) is for each firm j, the full marginal product of factor k in the production of good 1. The first term is the direct effect of factor k on good 1, and the second term is the indirect effect on good 1 acting through the reduction of exposure to z_{j1}.

Suppose that X_k is marketed competitively so that $q_k = p_k$. The consumer sets q_k/q_1 equal to the LHS of Eqn (8.27), and each producer sets p_k/p_1 equal to the RHS of Eqn (8.27). Therefore, competitive markets optimally allocate each affected firm's use of the defensive factor X_k.[12]

The answers, then, to the second and third questions above in this model are

1. A mixed strategy of taxing a pollutant at its source and using defensive measures to reduce exposure to the pollutant is pareto optimal. Also, the availability of defensive measures lowers the required Pigovian tax at the source because it reduces the other firms' exposure to the pollutant.
2. The government should not subsidize the use of defensive measures, providing the pollutant has been taxed (optimally) at its source. The intuition is that the full value of using a defensive measure is internal to each of the affected firms and is therefore a purely private good.

Equalizing Marginal Costs in Reducing Pollution

Another important result follows immediately from Eqns (8.26) and (8.27): When there is more than one strategy for reducing pollution, all strategies should be employed such that the MCs of each strategy are equal. Equalizing MCs across strategies ensures that any given reduction of pollution exposure is achieved at a minimum total cost to society. This result is analogous to the standard competitive result of equalizing MCs across firms in an industry to minimize the total costs of supplying a good or service.

The equal-marginal-cost interpretation comes from rearranging the profit-maximizing conditions for the firms. The tax raises firm 1's MC of using z_1 from zero to

$$t_z \Big/ \left(\frac{\frac{\partial f^l}{\partial z_l}}{\frac{\partial f^l}{\partial X_{ll}}}\right) = q_l \tag{8.28}$$

On the LHS, t_z is the price of z_1, the denominator is the marginal product of z_1 in the production of X_{11}, and the ratio is the MC of using z_1 expressed in terms of X_{11}. The MC of X_{11} equals its price q_1 given competitive markets. Also,

$$\left(\frac{\frac{\partial f^l}{\partial z_l}}{\frac{\partial f^l}{\partial X_{ll}}}\right) = -\sum \frac{\frac{\partial f^j}{\partial z_{jl}}}{\frac{\partial fj}{\partial X_{jl}}} \tag{8.29}$$

Therefore,

$$t_z \Big/ -\sum_{j=2}^{J} \frac{\frac{\partial f}{\partial z_{jl}}}{\frac{\partial fj}{\partial X_{jl}}} = q_l \tag{8.30}$$

The LHS equals the social marginal external cost of z_1 on the other firms, or the social MB of reducing z_1. In other words, the tax generates the standard MB = MC result for reducing pollution.

The MC of X_{jk} for firm j is

$$q_k \Big/ \left(\frac{\frac{\partial f^j}{\partial X_{jk}}}{\frac{\partial f^j}{\partial X_{jl}}} - \frac{\frac{\partial f^j}{\partial z_{jl}}}{\frac{\partial f^j}{\partial X_{jl}}}\frac{\partial z_{jl}}{\partial X_{jk}}\right) = q_1 \quad j = 1, \ldots, J \tag{8.31}$$

the ratio of the price of X_{jk} to the full marginal product of X_{jk}.

Notice that the MC to firm 1 of producing z_1 given the tax t_z and the MCs of using X_k are both equal to q_1. Therefore, the MCs of the tax and defensive strategies are equal, as required for reducing z_1 at minimum total cost to society.

Additional Complicating Issues

Our simple model ignores a number of complicating issues that are likely to be important in practice:

1. Defensive measures as externalities—In some instances, the defensive measures by affected agents may themselves generate an externality and require a Pigovian tax or subsidy on that account. A common example is

12. One interesting variation of the model would be an assumption that the use of X_{1k} by firm 1 benefits the other firms as well. Since firm 1 generates z_1, its attempts to reduce its own exposure could also reduce the amount of z_1 emitted from its site of production. This might be realistic for some kinds of pollution. Under this assumption, firm one is exposed to $z_{11} = z_1 - h^1(X_{1k})$, and the other $J-1$ firms are exposed to $Z_{J1} = z_{11} - h^j(X_{jk})$. X_{1k} thus gives rise to an externality just as z_1 does. The interested reader can verify that competitive marketing of X_{1k} would still be pareto optimal providing the government sets the optimal tax t_z on the production of z_1. The optimal tax not only guides firm 1 to the optimal production of z_1. It also causes the firm to incorporate correctly the external effect of X_{1k} when making its private decision regarding the use of X_{1k}. The intuition is that using X_{1k} reduces the firm's tax revenues (among its effects) by reducing its effective emission, z_{11}. Given t_z, the revenue reduction just equals the marginal external benefits to the other firms from the reduction in z_{11}.

vaccinations against diseases. People who receive vaccinations not only protect themselves from the disease, but also reduce the likelihood that others who are not vaccinated will contract the disease. This may explain why governments frequently subsidize vaccinations (Brito et al., 1991). Our model could incorporate this feature by having the h^j be a function of each firm's X_{jk}, either individually or in the aggregate. Pareto optimality would then require a Pigovian subsidy on each firm's use of X_{jk}. This will be left as an exercise for the interested reader.

2. Defensive measures as a medium for externality—Suppose that the effectiveness of the defensive measures depends on the level of the pollutant. We could represent this in our model as

$$z_{j1} = z_1 - h^j(X_{jk}, z_1)$$

This dependence generates two externality channels for z_1, one related to the damages inflicted on the other firms by the production of z_1 and the other related to the firms' costs in defending themselves. The second channel implies that the MCs of using X_{jk} depend on the level of z_1, which may often be the case. If so, then the MCs of defensive measures are more likely to be inversely rather than directly related to the level of the pollutant. A common pattern is that the first units of a pollutant can be removed relatively inexpensively, but then the costs of removal rise sharply once the pollutant reaches some low level. The inverse relationship can lead to nonconvexities and multiple equilibria for the abatement technology.

3. Very inexpensive defensive measures—Suppose the MCs of the defensive measures are so low that they can remove all exposure to the pollutant at an MC below the MC of reducing any pollution at its source. The laundry-factory example comes to mind. The pareto-optimal solution is to employ only defensive measures and not have any reduction of the pollutant at its source. Even so, the government should set a per-unit pollution tax on the polluters equal to the MC of the defensive measures to prevent unwanted entry into the polluting industry. The reason for maintaining a pollution tax in this case is as follows: Since the tax, by definition, is less than the MC of any abatement by the source, the polluter has no incentive to reduce the pollutant to save tax revenues. There is still no reduction of the pollutant at the source, as required for pareto optimality. At the same time, the pollutant does require costly defensive measures, and the tax acts as an entry fee into the industry. It discourages the entry of more polluters in the long run whose pollution would require still more defensive measures. In the laundry-factory example, the government does not want to encourage the building of more smoke-belching factories that might place still more moving costs on the laundries.[13]

4. Economies of scale—Our model so far implies that the affected firms and individuals should undertake defensive measures on their own. Large defensive measures undertaken or subsidized by the government do make sense, however, if they exhibit significant economies of scale. This is true even if pollution is taxed at its source. Economies of scale is an entirely separate issue from the externality point and will be analyzed in Chapter 9. The crux of the matter is that each agent affected by a pollutant should not be undertaking defensive measures at high MCs when a single large effort can be operated at much lower MCs to provide the protection. In fact, waste treatment of industrial water pollution does exhibit substantial scale economies.

The complicating factors 1 and 4 are most closely related to defensive strategies for global warming. For example, planting trees, whether in Brazil or elsewhere, generates external economies (diseconomies) to any individual or firm that is hurt (benefits) from global warming. And the plantings would have to be done on a massive scale. Given that the external effects of the plantings are global in scope, an international effort financed in part by all countries is the obvious way to proceed. Seeding the atmosphere, if that turns out to be viable, is also no doubt subject to economics of scale. Similarly, building seawalls to protect coastal cities calls for a joint effort. About all individuals and firms can do on their own is to move away from the coasts, which might be quite expensive.

LONG-LIVED EXTERNALITIES

Global warming raises one final issue that is partly philosophical and partly economic in nature. The GHGs remain in the atmosphere trapping heat for an incredibly long period of time. The Environmental Protection Agency compares the heat-trapping effect of each of the GHG 100 years out. CO_2 never dissipates. The gas continuously circulates between land, atmosphere, and ocean, disappearing only when it is absorbed into sediments at the bottom of the oceans. The point is that GHG emissions today generate costs across a large number of future generations.

This gives rise to the following tension. Economists typically assume that the goal of public policy over the long run is to maximize the sum of the utilities of each population cohort over time discounted to present value.

$$W = \sum_{t=0}^{\infty} U(C_t)\left(\frac{1}{1+\rho}\right)^t \tag{8.32}$$

13. These last two points, along with additional analysis of defensive strategies, are contained in Oates (1983).

The discount rate ρ applied to the different cohorts is called the social rate of time preference. ρ is not the same as the discount rate that individuals use to discount the future in their intertemporal budget constraints. That discount rate represents the net of tax return over time on the individuals' saving. Instead, ρ represents society's judgment about the appropriate worth of future cohorts relative to the current generation. The tension is that selecting a fairly high rate of time preference, even one that is just a couple of percentage points, discounts heavily the utilities of cohorts living one hundred years and more from now. Their utilities count for very little. Consequently, the costs of global warming borne by people three or four generations out and beyond have little effect on social welfare today, even if the costs they bear are substantial. (The current value of $100 of cost from global warming incurred 100 years from now is $1.98 discounted at 4%, $5.20 discounted at 3%, and $13.80 at 2%.) The implication is that current generations should not bear large costs today to reduce global warming because the costs of the global warming will be borne far in the future. Conversely, selecting a very low rate of time preference, say one near zero, on the grounds that the utilities of future cohorts should matter very much, can generate enormous costs of global warming because the costs keep coming more or less indefinitely, and they continue to matter just about as much as if they were incurred today. The implication is that the current cohorts should bear large costs today to reduce global warming for the benefit of generations in the distant future.

The tension between the current and future generations regarding global warming surfaced in 2006. The British government engaged economist Sir Nicholas Stern to head a commission to study the effects of GHG emissions and make recommendations about how much they should be reduced. The Stern commission issued its report, entitled *Stern Review on the Economics of Climate Change*, in November 2006. Its conclusions were quite dramatic:

"[T]he Review estimates that if we don't act, the overall costs and risks of climate change will be equivalent to losing at least 5% of global GDP each year now and forever. If a wider range of risks and impacts is taken into account, the estimates of damage could rise to 20% of GDP or more…Our actions now and over the coming decades could create risks…on a scale similar to those associated with the great wars and the economic depression of the first half of the 20th century" (p. xv). The Review recommended a price of carbon of $350 per ton, equal to its estimate of the social MC of carbon emissions in 2005 prices with the current no policy regime, roughly equivalent to $85 per ton of carbon dioxide emitted (p. 344).[14]

William Nordhaus wrote a critique of the *Stern Review* recommendations for the *Journal of Economic Literature* that was highly skeptical of the Review's recommendations. Nordhaus had long been involved with models of climate change developed at Yale University since 1975 called the Yale/DICE/RICE models.[15,16] The 2007 version of the DICE model, DICE-2007, reaches very different recommendations from those of the *Stern Review*. It calls for a harmonized carbon price phased in over time: $35 per ton of carbon emitted in 2015, rising to $85 per ton by 2050, and to $206 per ton by 2100, all expressed in 2005 prices. According to Nordhaus, the difference between the DICE-2007 and the *Stern Review* recommendations turn almost entirely on their different choices for the social rate of time preference.

Nordhaus notes that the centerpiece of the DICE and *Stern Review* models is the Ramsey–Koopmans–Cass model of optimal economic growth, which has become a workhorse model in macroeconomic analysis. It is essentially the original Solow growth model with saving and consumption determined endogenously rather than being fixed, as Solow assumed. Its objective function is social welfare as defined by Eqn (8.32) above, along with the simplification that $U(C_t) = \frac{1}{(1-\alpha)} c(t)^{(1-\alpha)}, 0 \leq \alpha \leq \infty$. This utility function, the same as used by Atkinson in his analysis of inequality (see Chapter 4), assumes a constant elasticity of the marginal utility of consumption, equal to $-\alpha$. Maximization of the social welfare function under the assumptions that population is constant and consumption per capita is growing at a constant rate g because of technical change leads to the following steady state equation:

$$r = \rho + \alpha g \tag{8.33}$$

where r is the real return to capital in equilibrium, equal to the marginal product of capital. Equation (8.33) is called the Ramsey equation.

Nordhaus centers his critique of the *Stern Review* on the Ramsey equation. The *Stern Review* argues that society should treat all generations equally, implying that the social rate of time preference, ρ, should equal zero. They set $\rho = 0.1$ to allow for some probability of future extinction, which is essentially the same zero. They also assumed $\alpha = 1$, that utility in each generation is the log of consumption, and that the rate of growth of consumption per capita is 1.3% per year.

Nordhaus finds this set of assumptions to be questionable in two respects. First, the near-zero rate of time preference implies that all future costs of global warming are essentially given the same weight as if they occurred today, no matter how far in the future they may be. In the *Stern*

14. Stern (2007). The quotation appears in the Executive Summary, p. xv and the recommended carbon price on p. 364. The choice of words in the quotation is by Nordhaus (2007).

15. DICE stands for Dynamic Integrated Model of Climate and the Economy.
16. RICE stands for Region Integrated Model of Climate and the Economy.

Review model, more than half of the costs of global warming that are seen to occur "now and forever" occur after 2800. In Nordhaus's view, setting the rate of time preference to zero gives far too much weight to distant future generations, especially given the huge uncertainties about what consumer preferences will be even 100 years from now. The notion that all generations should be treated equally sounds nice in principle, but the weight it gives to all future generations is too high for it to be compelling. To drive this point home, Nordhaus imagines the following "wrinkle experiment." Suppose climate scientists discover a wrinkle effect in the atmosphere that will lead to costs equal to 0.1% of consumption per capita starting in 2200, two centuries from now. On the assumptions used in the *Stern Review*, the current generation should sacrifice 56% of 1 year's world consumption today to avoid the wrinkle. But with consumption growing at 1.3% per year, and average current consumption per capita approximately equal to $10,000, average consumption per capita in 2200 will be $130,000. Therefore, under the *Stern Review* assumptions, the current generation should reduce its average consumption from $10,000 to $4400 to prevent a decline in consumption from $130,000 to $129,870 starting in 2200 and continuing forever after. This trade-off to avoid such a tiny cost to much richer future generations is simply unconvincing, in Nordhaus's view.

A second criticism is that the parameters chosen by the *Stern Review* imply a real return to capital of 1.4% ($1.4 = 0.1 + 1(1.3)$), whereas the actual return to capita is on the order of 5.5%. The *Stern Review* therefore assumes a much higher saving rate and capital stock (lower marginal product of capital) than actually exists. Nordhaus believes that projections of the costs of global warming should pay attention to actual market rates of interest. Therefore, when projecting the costs of global warming and the harmonized carbon price that they imply in his DICE-2007 model, he assumes: $g = 2\%$, which is the average rate of growth of consumption predicted by the DICE-2007 model over the next century, a rate of time preference $\rho = 1.5\%$, and an elasticity of the marginal utility of consumption $\alpha = 2$. These parameters are consistent with the real return to capital of 5.5% ($5.5 = 1.5 + 2(2)$). These are the parameters that lead to the phased set of carbon prices reported above. If he runs the DICE-2007 model with a $\rho = 0.1$ and $\alpha = 1$, he generates an increase in the price of carbon to $360, as did the *Stern Review*, highlighting that the *Stern Review* results and recommendations are almost entirely due to the assumptions on these two parameters, particularly the near-zero rate of time preference.

By using a set of parameters that far underestimate the marginal return to capital, the *Stern Review* recommendations ignore the current high productivity of capital in producing consumption goods. As a result, they give too much weight to capital devoted to reducing carbon emissions. The better strategy is to exploit the high return to capital in the near term to increase the growth in consumption per capita, and only later sharply raise the price of carbon in stages to reduce the costs of global warming. This phased-in strategy allows all generations to be better off. In fact, under Nordhaus's parameters, the immediate large increase in the price of carbon recommended by the *Stern Review* makes future generations worse off because it reduces their consumption per capita by not exploiting the high return to capital in producing consumption goods now and in the near future.

In summary, the Nordhaus critique highlights the care that has to be taken when modeling events that have consequences far into the future. It is not clear what the right model is, but he recommends trying a number of different assumptions to assess the possibilities of very-long-run projections. He does not find the recommendations of the *Stern Review* at all convincing.[17]

REFERENCES

Buchner, B., Carraro, C., Cersosimo, I. and Marchiori, C., "Back to Kyoto? US Participation and the Linkage between R&D and Climate Cooperation," Discussion Paper No. 3299, Centre for Economic Policy Research, April 2002.

Brito, D., Sheshinski, E., Intriligator, M., June 1991. Externalities and compulsory vaccinations. Journal of Public Economics 45 (1), 69–90

Cropper, M., Oates, W., June 1992. Environmental economics: a survey. Journal of Economic Literature XXX (part IV), 675–740

Mills, E., 1967. Economic incentives in air-pollution control. In: Goldman, M. (Ed.), Controlling Pollution: The Economics of a Cleaner America. Prentice-Hall, Englewood Cliffs, NJ

Nordhaus, W., September 2007. A review of the stern review on the economics of climate change. Journal of Economic Literature 45 (3), 686–702

Nordhaus, W., December 7, 2005. Life after Kyoto: Alternatives to Approaches to Global Warming Policies. Yale University and the NBER

Oates, W., 1983. The regulation of externalities: efficient behavior by sources and victims. Public Finance 38 (3), 362–375

Stern, N., 2007. The Economics of Climate Change: The Stern Review. Cambridge University Press, Cambridge and New York

United Nations Framework on Climate Control. Web sites: http://unfccc.int/kyoto_protocol/items/2830.php; http://unfccc.int/resource/docs/2009/cop15/eng/l07.pdf; http://unfccc.int/home/items/5264.php; http://unfccc.int/home/items/5265.php.

Weitzman, M., 1974. Prices vs. Quantities. Review of Economic Studies 41 (4), 477–492

U.S. Environmental Protection Agency. Web site: epa.gov/climatechange/science/indicators/ghg/index.html.

17. Stern (2007). The quotation appears in the Executive Summary, p. xv and the recommended carbon price on p. 364. The choice of words in the quotation is by Nordhaus (2007). Nordhaus compares the *Stern Review* and his preferred parameters for the Ramsey equation on p. 694. His phased-in estimates for the price of carbon in the DICE model are on p. 698, and the DICE result of a $360 price of carbon using the *Stern Review* Ramsey equations parameters is on p. 698. His wrinkle experiment is on p. 696, as is his statement that half of the damages in the *Stern Review* occur after 2800.

Chapter 9

The Theory of Decreasing Cost Production

Chapter Outline

Production of some goods and services exhibits significant decreasing cost or increasing returns to scale, meaning that unit or average cost for an individual firm continues to decline up to a substantial proportion of total market demand. Whenever this occurs, government intervention is almost certainly required to ensure a social welfare optimum.

Public sector economics has traditionally concerned itself only with the most extreme example of decreasing cost, in which a single firm's average cost declines all the way to total market demand. This is referred to as a *natural monopoly*, because a single firm can supply the entire industry output most cheaply. The problems arising in less-extreme instances of decreasing cost production that lead to oligopolistic market structures have traditionally been covered in courses on industrial organization. In keeping with this tradition, this chapter analyzes only the natural monopoly, so that "decreasing cost" means decreasing unit cost to total market demand.

Decreasing cost industries are not at all rare. They typically occur in the production of services rather than products—in particular, services that require relatively large setup costs, after which large numbers of users can be served at relatively low marginal cost. The combination of high startup costs and low marginal cost causes unit cost to decline even as the number of users becomes large. Services having these attributes include many forms of transportation, especially highways, bridges, tunnels, and rail transit; the so-called public utilities, such as telephone service, electricity, water supply, and sewage; first-class mail delivery; some recreational facilities such as beaches and parks; some forms of entertainment such as radio, television, and the commercial use of songs; telecommunications generally, and the software, data, and other services available on the Internet.

The entertainment examples and the Internet are among the purest instances of decreasing cost services. Think of the viewing of television programs. Considerable resources are required to produce, transmit, and receive any one television program. But once the program is produced and the transmitting and receiving facilities (televisions) are in place, the cost of another viewer turning on his television set is essentially zero, no matter how many people are

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00009-1

watching. Because the number of viewers is the relevant unit of output, average cost decreases continuously as the number of viewers increases. The same properties apply as well to the Internet. Software and data services are often sold commercially on CD-ROMs, with accompanying manuals, so that their producers can earn a profit, but the software and data could be provided over the Internet (and sometimes are), where they would simply be downloaded by whoever wants them. Clearly the lowest marginal cost of providing software and data is virtually zero. They are exactly like songs in this respect. All the cost and effort are in the creation of the work (i.e., the setup cost).

Marginal costs are greater than zero for each of the other services listed above, but they are nonetheless relatively unimportant compared with the fixed costs of establishing the services. For example, the decreasing cost property of the public utilities or first-class mail arises in the distribution of the services, the need to set up a network of telephone and electric lines (or satellites), water pipes, or post offices to reach every household. Similarly, a large percentage of the costs of providing and maintaining highways, bridges, rail transit, beaches, and parks are essentially fixed costs, unrelated to the number of users.[1]

Decreasing cost production requires government intervention in a market economy for the simple reason that it is totally incompatible with a competitive market structure. Decreasing cost industries cannot possibly have a competitive structure, with large numbers of price-taking firms. Moreover, even if the competitive structure were possible, it would not be desirable. In order to capture the benefits of increasing returns production, the entire output must be produced by a single firm. This is why decreasing cost industries are referred to as "natural monopolies," and why they necessarily violate the technological assumption of well-behaved production required for a well-functioning competitive market system.

This chapter explores the pareto-optimal conditions for efficient production with decreasing cost industries. It shows why the competitive market system cannot achieve them and considers the pricing and investment rules implied by the efficiency conditions. The pricing rules imitate standard competitive principles and are therefore relatively straightforward. The investment rules are far from standard, however. Investment in decreasing cost industries has a lumpy, all-or-none quality to it that is absent in the usual marginal investment analysis applied to the small competitive firm. Also, profitability is not necessarily a reliable investment guideline for decreasing cost services, and there may not be any other practicable criteria to determine whether an investment in these services is worthwhile. As a result, investment decisions for these industries are frequently among the more difficult decisions the government has to make, even under the simplifying assumptions of first-best theory.

The chapter concludes with a discussion of actual U.S. policy with respect to the decreasing cost services. Governments in the United States have not embraced the first-best price and investment decision rules for decreasing cost services. They tend to favor some form of average cost pricing and the standard private sector profitability criteria for investment decisions, neither of which is pareto optimal. The policy discussion speculates on the popularity of these policies and analyzes their properties relative to the first-best decision rules.

DECREASING COST IN GENERAL EQUILIBRIUM ANALYSIS

The problems caused by a decreasing cost natural monopoly are directly related to the particular form of its production function, nothing more. Therefore, the general equilibrium model required to analyze decreasing cost production can be extremely simple. There is no need to model explicitly the interrelationships among consumers or producers, unlike in the analysis of externalities. Also, the model can exploit the dual dichotomies of all first-best models. The demand side of the model can be adequately represented by a single consumer. Were the model to include many consumers and a social welfare function, the first-order conditions would merely reproduce the interpersonal equity conditions and the pareto-optimal consumption conditions of the full model in Chapter 2. Assuming optimal redistributions implies a one-consumer equivalent economy. Similarly, positing many "well-behaved" firms would reproduce the standard pareto-optimal production conditions for those firms. Hence, a single producer is also sufficient so long as its production exhibits increasing returns to scale. Finally, there is no need to specify N goods and factors. A one-good, one-factor economy is sufficient to represent increasing returns or decreasing cost production. Consequently, a general equilibrium model consisting of one person with one source of (decreasing cost) production at which a single output is produced by means of a single input is sufficiently general to capture both the nature of the decreasing cost problem and the decision rules necessary to ensure full pareto optimality. Keeping the model this simple has an additional advantage of permitting a two-dimensional geometric analysis, a welcome relief from the notational complexities of the various externality models. Therefore, let us assume

1. A single consumer with utility function:

$$U = U(X, L) \tag{9.1}$$

1. Marginal costs do rise considerably for these services when they become congested, but congestion is an example of an externality. It is unrelated to the phenomenon of significant scale economies or decreasing cost.

where

$X =$ the single output.
$L =$ labor, the only factor of production.

By the usual convention, L enters $U(\ )$ with a negative sign (leisure is the good). The indifference curves corresponding to $U(X, L)$ are represented in Fig. 9.1.

2. A single firm produces X according to the production function:

$$X = f(L) \tag{9.2}$$

with $\partial f/\partial L \equiv f' > 0$, $\partial^2 f/\partial L^2 \equiv f'' > 0$, and $f(0) = 0$. The increasing returns production function is represented in Fig. 9.2.

$f(L)$ is a homogeneous function exhibiting increasing returns to scale. As such, average cost continuously decreases. This follows because $Lf' > f$ from Euler's equation on homogeneous functions; thus, $f' > f/L$ (the slope of f at any point is greater than the slope of a ray from the origin), and

$$\text{AC} = P_L/f/L > P_L/f' = \text{MC} \tag{9.3}$$

everywhere. Average cost (AC) and marginal cost (MC) decrease continuously, with MC < AC.[2]

Market clearance automatically holds in this model since it is understood that the consumer supplies all the labor used by the firm and consumes all the output produced by the firm. In a market context, the consumer also receives (pays) all pure economic profits (losses) arising from production at prices P_X and P_L, as well as earning all the labor income.

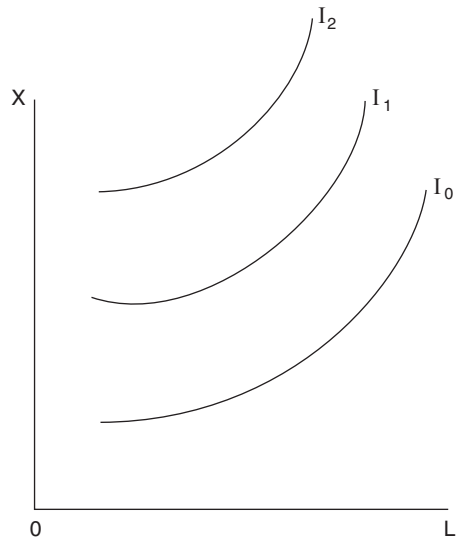

FIGURE 9.1

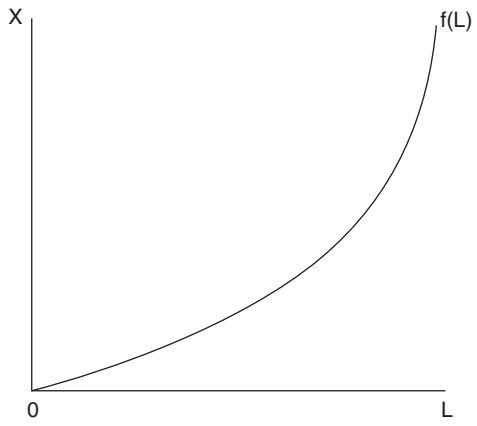

FIGURE 9.2

The Pareto-Optimal Conditions

Society's problem is the standard one:

$$\max_{(X,L)} U(X, L)$$
$$s, t.\ X = f(L)$$

To derive the pareto-optimal conditions, substitute the production function into the utility function and solve the unconstrained maximum:

$$U_X f' + U_L = 0$$

The first-order conditions are

$$U_X f' + U_L = 0 \tag{9.4}$$

or

$$U_X f' = -U_L \tag{9.5}$$

Condition (9.5) gives the standard result that labor should be used to produce X until the marginal utility of X just equals the marginal disutility of further work.

The second-order conditions cannot be ignored with increasing returns-to-scale production, as both the indifference curves and the production function have the same general curvature. To ensure that condition (9.5) represents a utility maximum, the derivative of Eqn (9.5) with respect to L must be negative. Thus,

$$\frac{\mathrm{d}\left(f' + \frac{U_L}{U_X}\right)}{\mathrm{d}L} < 0 \tag{9.6}$$

or

$$\frac{\mathrm{d}f'}{\mathrm{d}L} + \frac{\mathrm{d}\left(f' + \frac{U_L}{U_X}\right)}{\mathrm{d}L} < 0 \quad (\text{Note; } U_L < 0) \tag{9.7}$$

2. The appendix demonstrates the relationship between increasing returns to scale and decreasing cost for the general case of more than one factor of production.

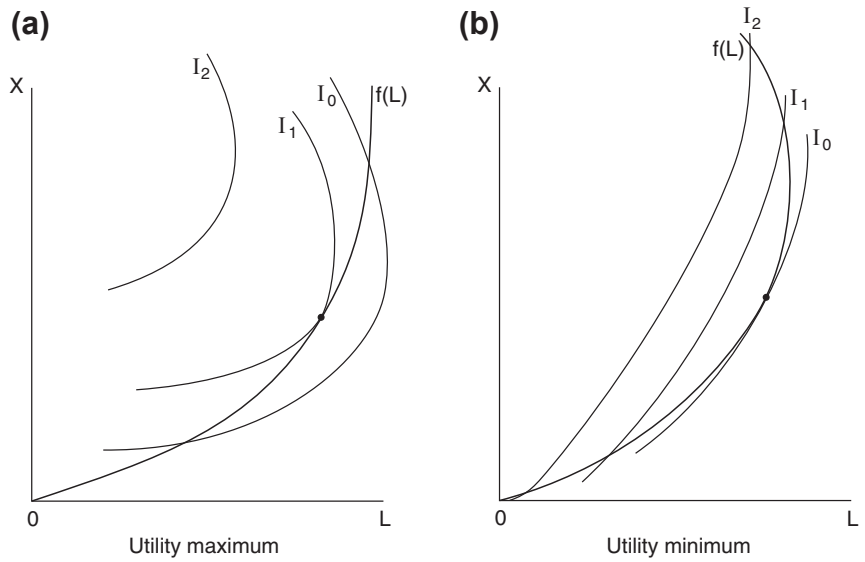

FIGURE 9.3

Equation (9.7) implies that the curvature of the indifference curves for X and L must be greater than the curvature of the production function. If the reverse is true, Eqn (9.5) represents a utility *minimum* along the production frontier. Refer to panels (a) and (b) in Fig. 9.3.

Decreasing Cost and Competitive Markets

A competitive industry cannot achieve condition (9.5) even if the second-order conditions for a welfare maximum are satisfied. A price-taking competitive firm would solve the following problem:

$$\max_{(L)} P_X f(L) - P_L L$$

The first-order conditions are

$$P_X f' - P_L = 0 \tag{9.8}$$

Alternatively,

$$P_X = P_L/f' = \mathrm{MC}_X \tag{9.9}$$

Conditions (9.8) and (9.9) are the familiar results that the competitive firm hires labor such that the value of the labor's marginal product equals the price of the labor or, equivalently, supplies X such that price equals marginal cost.

On the surface, this result would appear to satisfy the full pareto-optimal conditions. The consumer maximizes utility by equating

$$\frac{U_X}{P_X} = \frac{-U_L}{P_L} \tag{9.10}$$

or

$$P_X = -\frac{U_X}{U_L} \cdot P_L \tag{9.11}$$

Substituting for P_X in Eqn (9.9) yields

$$-\frac{U_X}{U_L} \cdot P_L \cdot = P_L/f' \tag{9.12}$$

or

$$U_X f' = -U_L \tag{9.13}$$

the pareto-optimal condition (9.5).

Perfect competition cannot be pareto-optimal, however, because setting the price of labor equal to the value of its marginal product is *not* profit maximizing for a decreasing cost firm. Since $P_X f'' > 0$, the second-order conditions for a maximum fail to hold. Thus, Eqn (9.9) is the *profit-minimizing* condition.

The perfectly competitive, decreasing marginal cost firm would maximize profits by increasing output indefinitely, as depicted in Fig. 9.4. Since marginal cost P_L/f' declines continuously,[3] the firm loses on every unit up to X_0, but that is the output given by Eqn (9.9). Marginal profits begin at $X_0 + 1$ and increase indefinitely as X increases. The industry would finally consist of a single firm producing the entire market demand.

Competition and decreasing costs are incompatible in another sense. Even if the government were able to subsidize the losses suffered by each firm producing at X in

3. $d(P_L/f') = -P_L f''/(f')^2 < 0$. Strictly speaking, marginal costs could be constant or rising because the term *decreasing cost* refers to decreasing unit cost. But, since $\mathrm{MC} < \mathrm{AC}$ when AC is declining, any $P = \mathrm{MC}$ "equilibrium" implies losses for the competitive firm.

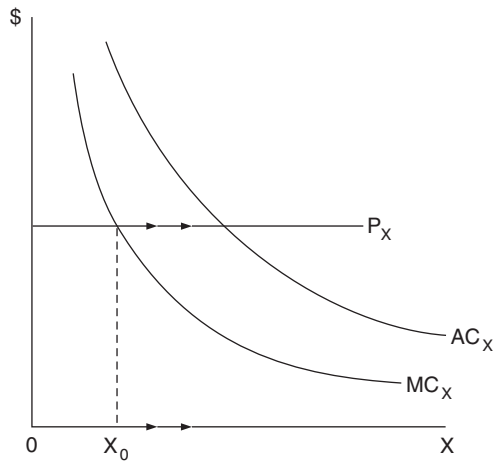

FIGURE 9.4

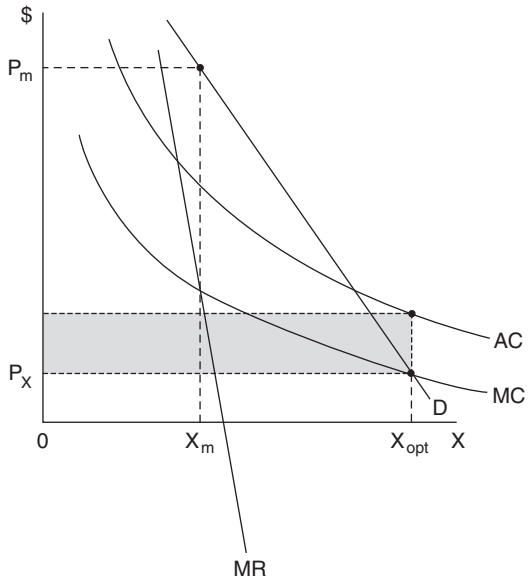

FIGURE 9.5

Fig. 9.4, society would not want this solution. One firm has to become the entire industry to exploit the scale economies and minimize production costs, resulting in a natural monopoly. Figure 9.5 is the relevant diagram, with a single firm facing the entire market demand curve.

The Optimal Pricing Rule

Efficiency requires that pareto-optimal condition (9.5) hold; it is the first-order condition for a pareto optimum. But, this implies that the monopoly firm must set price equal to marginal cost. Only if $P_X = P_L/f' = \text{MC}_X$ will $U_X f' = -U_L$. Referring to Fig. 9.5, the monopolist must produce at X_{opt}, the point at which the market demand and marginal cost curves intersect.

There are two problems with the pareto-optimal solution, however. First, an unregulated, profit-maximizing monopoly would not produce X_{opt}; instead, it would choose X_{m}, at which MR = MC, and set $P_{\text{m}} > P_X$. Thus, the government must either force the monopoly to select (P_X, X_{opt}) through regulation or operate the industry itself. Second, a regulated monopoly (or the government) makes perpetual losses if forced to produce at $P_X = \text{MC}$, with AC > MC. Since the monopoly must cover its full costs in the long run, it must be subsidized by an amount equal to $X_{\text{opt}} \cdot (\text{AC} - \text{MC})$, the shaded area in Fig. 9.5. Moreover the subsidy must not generate inefficiencies elsewhere in the economy. It must be lump sum. In this simple economy, the consumer would transfer the required income to the firm. Such a transfer is possible, because the excess of the consumer's earned income over his or her expenditures, $P_L L - P_X X$, exactly matches the firm's deficit.

In a many-person economy, the transfers to decreasing cost firms become part of the lump-sum redistributions necessary to satisfy the interpersonal equity conditions. The only difference from the optimal lump-sum redistributions described in Chapter 2 is that the taxes collected must be sufficient to cover both the transfers to other individuals and the transfers to decreasing cost firms. Social marginal utilities are still equalized at the first-best optimum.

To summarize, the first-best pricing rule is $P = \text{MC}$, with a lump-sum subsidy to cover the deficit with marginal cost pricing.

The Optimal Investment Rules

When confronted with decreasing cost production, society must make a fundamental investment decision that does not present itself in "normal" industry situations. If the firm is forced to price optimally at $P = \text{MC}$, it cannot cover the opportunity costs of investing in that industry. The lump-sum subsidy does allow the firm to cover opportunity cost, but this in itself is not especially helpful since profit (loss) is not performing its customary function as a signal for investment. Society is, instead, presented with an all-or-none decision: Is providing the service at $P = \text{MC}$, with a subsidy to cover the operating deficit, preferable to not having the service at all? This is obviously quite different from the standard investment decision, in which the present value formula is used to determine the profitability of a marginal increment to the capital stock evaluated at current (and expected) market prices.

One can usefully distinguish two cases in analyzing the investment decision, which we shall refer to as the "easy" case and the "hard" case. The distinction between them

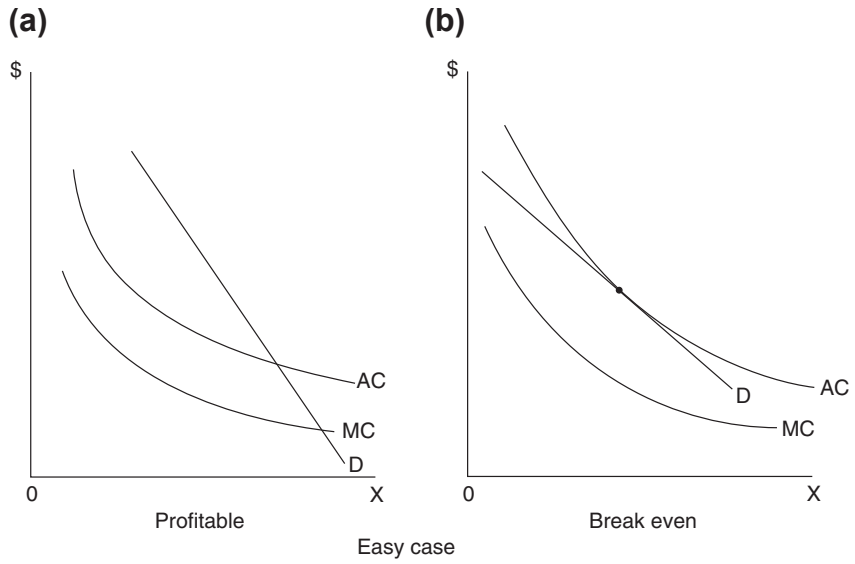

FIGURE 9.6

turns on whether a profit-maximizing monopolist could at least break even *if* it were allowed to do so.[4]

The Easy Case

Suppose we know that demand is sufficiently high so that a monopolist, if allowed to maximize profit, could at least break even by charging a single price for its product. In other words, the market demand curve is at least tangent to the firm's AC curve, as in the two panels of Fig. 9.6. This case would certainly apply to most public utilities; urban highways, bridges, and tunnels; television and radio; telecommunications including the Internet; and many recreational facilities.

We will show that if (at least) break-even production is possible, the service should be produced. Potential profitability is thus a sufficient condition for having the service. The monopolist must not be allowed to capture these potential profits, however. Price must be set at marginal cost, with a lump-sum subsidy to cover the resulting losses at that price.

The Hard Case

The ability to break even is not a necessary condition for providing the service, however. Suppose demand is everywhere below AC, as in Fig. 9.7, such that there is no single price at which a profit-maximizing monopolist could break even. Clearly, no private firm would be interested in this market unless heavily subsidized. Society may be interested in having the government provide the service, however, once again at the intersection of price and marginal cost. Then again, society may not be interested. Demand may be so low that the service is not worth having.

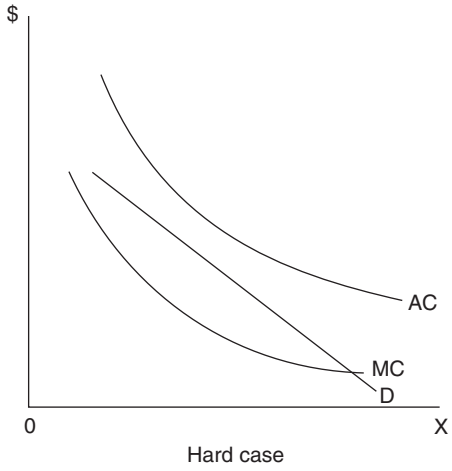

FIGURE 9.7

Notice that even potential profitability is not a useful investment guideline for the hard case. As we shall see, the necessary conditions involve willingness-to-pay criteria that do not have close market analogs.

The hard case is not merely a theoretical *curiosum*. Many, if not most, rural highways would certainly fall into this category, as well as a number of recreational facilities such as national parks in remote areas. Urban (heavy) rail transit systems also appear to be examples of the hard case. The

4. Because capital is suppressed in the simple geometric analysis, we are implicitly assuming that the capital stock is optimal for any given X; that is, AC and MC are minimum long-run costs. The all-or-none test is an additional question, asking whether first-best optimal production and pricing is preferred to having no service at all. Our analysis of the all-or-none test follows that of Diamond and McFadden (1974).

American Public Transportation Association used to publish revenue and cost data for the rail transit systems in the 11 U.S. cities that have them. In FY 2002, the last year we could find published data on the individual systems, not one system had revenues that covered even close to half of its operating costs, much less any of its capital costs. This is hardly for want of trying; cities often raise transit fares in an attempt to reduce their deficits. It seems quite likely that no fare on any of these systems would be able to cover their full costs (American Public Transportation Association, 2004).

The Sufficient Condition: The Easy Case

Let us first establish the sufficient condition for decreasing cost production.

Proposition: If a profit-maximizing monopolist can at least break even by charging a single price, then society should produce the good. Utility is maximized, however, by setting $P = \text{MC}$, and covering the resulting deficit with a lump-sum transfer.

A geometric proof will suffice, provided we use the consumer's indifference map and the production–possibilities frontier in (X, L)-space. The demand curve-average cost diagrams are useful for illustrating certain points but are illegitimate as a representation of general equilibrium. They rely on inappropriate measures of consumer's surplus.[5]

Break-even Production

The first step in the proof is to characterize break-even production in $(X–L)$-space. Refer to Fig. 9.8. Think of production as occurring somewhere along the ray 0R from the origin, say at point B. Suppose the monopolist were to set the relative prices P_L/P_X equal to the slope of 0R. Let the slope equal k. The monopolist would then just break even, since at any point along the ray:

$$X = P_L/P_X \cdot L \text{ or } P_X X = P_L L \qquad (9.14)$$

This example is *not* to suggest a monopolist would actually set relative prices equal to k, but only to depict the limiting, breakeven, $P = \text{AC}$ case.

The Price–Consumption Locus

The ray 0R through B, with relative prices $P_L/P_X = k$, also serves as a budget line for the consumer with no lump-sum taxes or transfers. Suppose the consumer happens to be in equilibrium at point B as shown in Fig. 9.9.

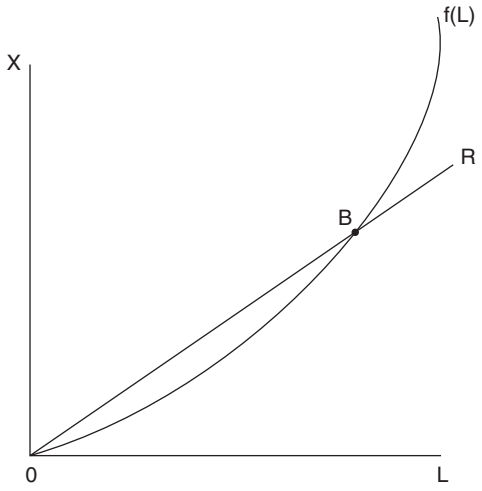

FIGURE 9.8

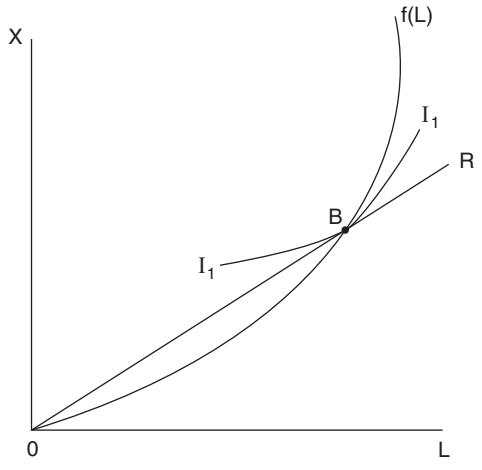

FIGURE 9.9

Then, point B represents an actual general equilibrium for the economy with prices such that production is breakeven.

We can trace out a price–consumption line for the consumer by varying the slope of the ray through the origin, as in Fig. 9.10, with P_L/P_X always equal to the slope. Every point on the price–consumption locus represents a potential general equilibrium for the economy, in which the consumer is in equilibrium and the firm is breaking even.

The only remaining question is whether a general equilibrium with break-even production is feasible. The answer is yes if the price–consumption locus intersects the production–possibilities frontier, such as at point B in Fig. 9.11. Furthermore, break-even production at point B is preferred to the origin, the point of zero production. This follows because the consumer is in equilibrium at B, so that the indifference curve on which point B lies is tangent to the ray through the origin. Therefore, B must lie above the

5. The geometric proofs of the necessary and sufficient conditions were demonstrated to us by Peter Diamond in his graduate Public Finance class at MIT. See also Diamond and McFadden (1974), especially for the hard case.

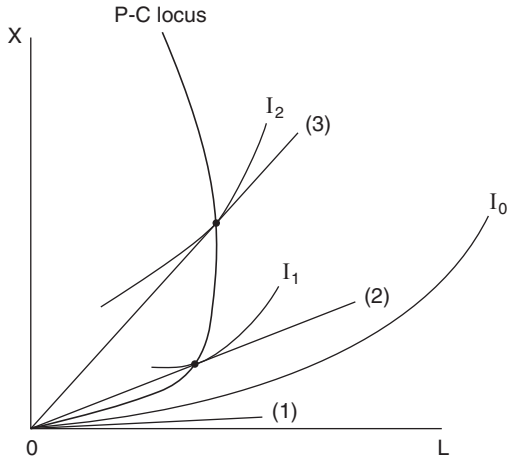

FIGURE 9.10

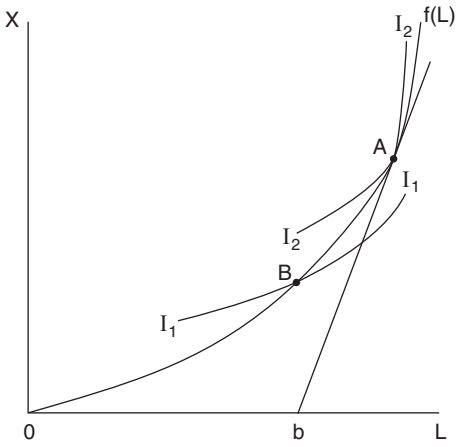

FIGURE 9.12

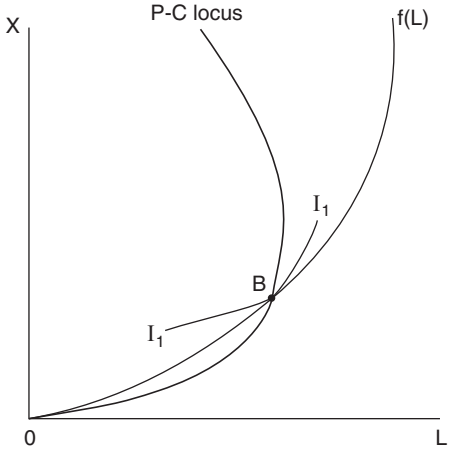

FIGURE 9.11

indifference curve that passes through the origin, I_0, establishing the first part of the proposition:

Feasibility of break-even production is a sufficient condition for having the service.

Production is not pareto optimal at B, however. The problem with point B is that P_X equals the average cost of producing X, not the marginal cost. The slope $k = X/L$ is the average product of labor at B. Therefore,

$$P_X = P_L/k = P_L/\mathrm{AP}_L = \mathrm{AC}_X \qquad (9.15)$$

The marginal product of labor at B is f', which is greater than k, the average product. Therefore

$$\mathrm{P}_L/f' = \mathrm{MC}_X < P_L/k = \mathrm{AC}_X = P_X \qquad (9.16)$$

in violation of the pareto-optimal condition (9.5).

The utility from the service is maximized by producing at point A in Fig. 9.12, at which point the production–possibilities frontier is just tangent to one of the consumer's indifference curves, I_2.

The remaining question is how to establish point A as a marketed general equilibrium. This is done by setting P_L/P_X equal to the slope of the production–possibilities frontier at A, so that $P_X = \mathrm{MC}_X \equiv P_L/f'$. This price ratio is the slope of a ray intersecting the L-axis at point b, not the origin. This ray can be a budget line for the consumer only if the consumer first surrenders b units of labor lump sum, so that

$$P_X X - PL \cdot L = -P_L b \qquad (9.17)$$

Similarly, with $P_X = \mathrm{MC}_X$ at A, the firm makes pure economic losses equal to $P_L b$, since the ray is also the firm's profit line. Therefore, with a lump-sum transfer of $P_L \cdot b$ from the consumer to the firm and with marginal cost pricing, the firm covers its full costs and the consumer attains the highest possible indifference curve. This is the price-with-transfer general equilibrium market solution that satisfies pareto-optimal condition (9.5), in line with the second part of the proposition.

The Necessary Condition: The Hard Case

The existence of a break-even production point, such as B in Fig. 9.11 implies the existence of a preferred pareto-optimal point such as A, but a pareto-optimal point A preferred to zero production does not imply B. Suppose the price–consumption line lies everywhere above $f(L)$ (except at the origin), as in Fig. 9.13. Since the price–consumption locus defines all possible break-even general equilibrium points, break-even production is not feasible. This corresponds, in (P_X,X)-space, to the situation in which the demand curve for X is everywhere below the AC curve.

Society may still prefer production at $P_X = \mathrm{MC}$ to not having the service, however. The all-or-none test turns on the position of the indifference curve passing through the

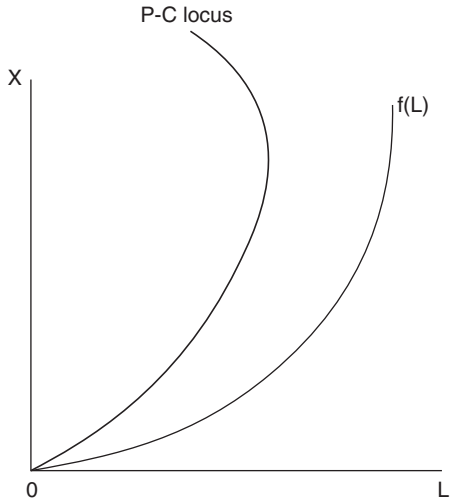

FIGURE 9.13

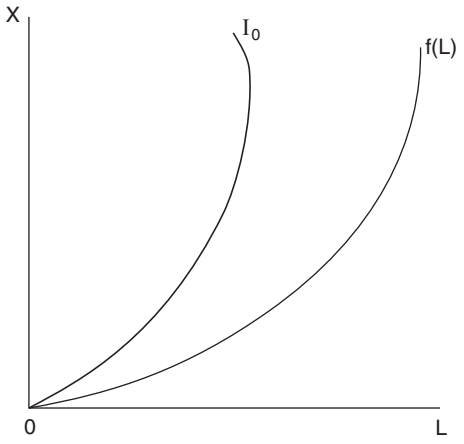

FIGURE 9.14

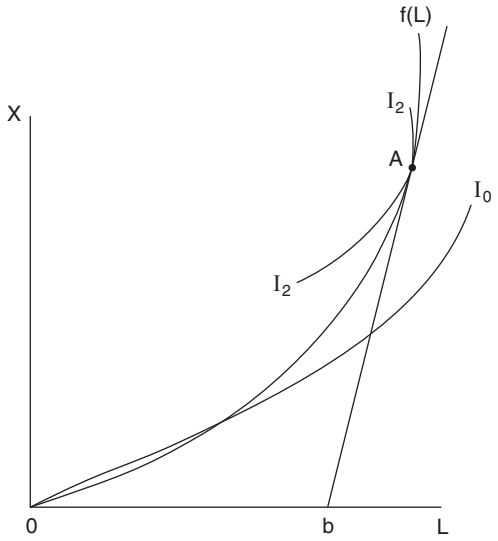

FIGURE 9.15

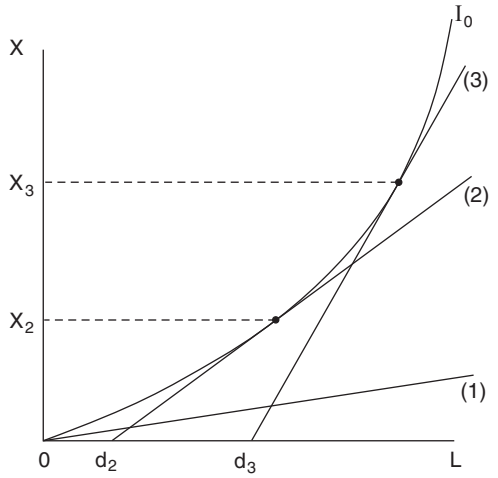

FIGURE 9.16

origin, I_0. If I_0 lies everywhere above the production–possibilities frontier as in Fig. 9.14, society should not produce X. The consumer prefers zero production to any of the feasible choices lying on or below $f(L)$.

If I_0 crosses $f(L)$ as in Fig. 9.15, then there exists a higher indifference curve tangent to the frontier, such as I_2 in Fig. 9.15. Society should produce at the tangency point A in Fig. 9.15, set $P_X = \text{MC}_X$, and transfer b units of labor lump sum from the consumer to the firm, exactly as in the easy case. This is the only way to satisfy pareto-optimal condition (9.5).

Notice that even potential profitability is no guideline in the hard case. Even if the firm were allowed to maximize profits, it could not do so much as break even by setting a single price. The government must rely instead on willingness-to-pay lump-sum income measures of welfare, such as Hicks' Compensating Variation (HCV), to determine whether production is worthwhile.

Consider the indifference curve through the origin I_0 in Fig. 9.16 and the lines (1), (2), and (3) tangent to I_0. The tangency lines (1), (2), and (3) represent budget lines for the consumer, in which:

1. P_X is decreasing such that the ratio of prices P_L/P_X equals the slope of the corresponding line.
2. The consumer first sacrifices (lump sum) 0, $0d_2$, and $0d_3$ units of labor, respectively, to remain on I_0.

For example, P_X is so high (relative to P_L, assumed constant) on budget line (1) that the consumer purchases no X. Consider line (3). Since utility is being held constant along I_0, the distance $0d_3$ can be interpreted as the lump-sum income (in terms of labor) the consumer is willing to pay for the opportunity to purchase X at the lower

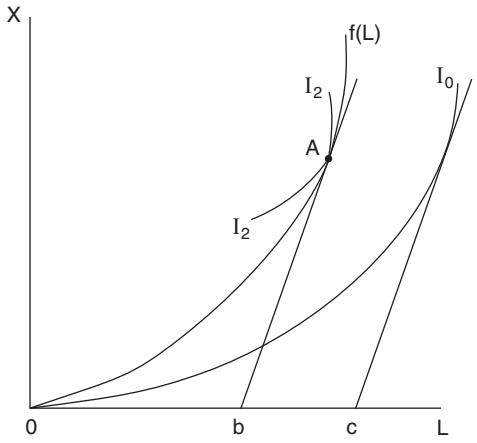

FIGURE 9.17

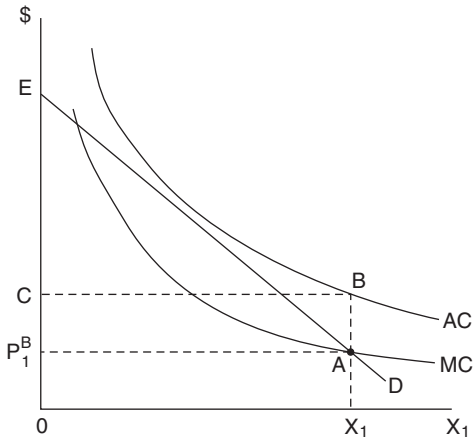

FIGURE 9.18

P_X (equal to P_L/slope of (3)). The income sacrificed keeps utility constant by exactly offsetting the utility gains resulting from the reduction in P_X. Hence, $0d_3$ is the HCV for this P_X, measured relative to a P_X so high that quantity demanded is zero. Also, X_3 is a point on the consumer's compensated demand curve for X, compensated to equal the utility level represented by I_0.

The HCV can then be compared with the actual amount of income the consumer must sacrifice for feasible decreasing cost production. Consider Fig. 9.17. At point A on I_2 with marginal cost pricing, the firm requires $0b$ units of L to break even. But, at the marginal cost price, the lump-sum amount the consumer is willing to pay is $0c$, the HCV measured along I_0. The consumer only has to pay $0b$ as a lump-sum transfer, so society should provide the service. In contrast, with I_0 everywhere above the $f(L)$ as in Fig. 9.14, the consumer would never be willing to sacrifice the lump-sum income required for the firm to cover its cost with marginal cost pricing. Hence, the service should not be provided.

To summarize, the necessary condition for providing the service is that the consumer's HCV, evaluated at $P_X = \text{MC}_X$ and the utility level with zero production, exceed the firm's deficit at the marginal cost price.

This test can also be represented in (P_X,X)-space, an interpretation worth analyzing because it appears in many sources, especially intermediate-level texts. Consider the market for X_1 as represented by the demand and cost curves in Fig. 9.18. The necessary condition is often stated as follows: If the area EP_1^B A exceeds the fixed-cost subsidy, area $CBAP_1^B$, then society should operate the service at X_1. This test is flawed because the measure of the consumer's benefit is Marshallian consumer surplus, which has no willingness-to-pay interpretation. But, if the demand curve is the compensated demand curve and only P_1 varies, then this test is equivalent to the necessary condition derived in (X,L)-space.

The Necessary Condition and the Compensated Demand Curve

To see this, suppose the government is considering whether or not to operate a new, decreasing cost industry at $P = \text{MC}$. Should the government decide not to produce, the consumer remains at the status quo (say A), with a utility level equal to U^{A}. Should the government produce, the consumer is in situation B, at a utility level of U^{B}. The question is whether $U^{\text{B}} \leq \geq U^{\text{A}}$.

The expenditure function provides the answer, since

$U^{\text{B}} > U^{\text{A}}$ iff $M(\vec{\mathbf{P}}; U^{\text{B}}) > M(\vec{\mathbf{P}}; U^{\text{A}})$ for any price vector $\vec{\mathbf{P}}$.

That is, for any price vector $\vec{\mathbf{P}}$, situation B costs more if it represents greater utility.[6]

To see how the expenditure function generates the all-or-none tests described above, begin by defining T^{A} as the lump-sum payment required of the consumer to support the equilibrium at A. Therefore,

$$M\left(\vec{\mathbf{P}}^{\text{A}}; U^{\text{A}}\right) = -T^{\text{A}} \tag{9.18}$$

or

$$\sum_i P_i^{\text{A}} X_i^{\text{A}}\Big|_{\text{comp}} = -T^{\text{A}} \tag{9.19}$$

Note that $-T^{\text{A}}$ could well be zero. It measures lump-sum income from taxes, transfers, fixed factors, and pure economic profits. With no decreasing cost production and all income earned, it is reasonable to assume $-T^{\text{A}} = 0$, although this is not required.

6. The sufficient condition follows directly from the assumption of non-satiation, which implies a positive marginal utility of income. The necessary condition follows from the property that well-behaved indifference curves cannot cross. Thus, given two market situations with identical prices, the one with higher income must generate higher utility.

Situation B with decreasing cost production requires an additional transfer from the consumer to the firms, thereby increasing the total payments by the consumer to T^B. Therefore,

$$M\left(\vec{\mathbf{P}}^B; U^B\right) = -T^B \tag{9.20}$$

To determine whether $U^B > U^A$, evaluate the expenditure functions at $\vec{\mathbf{P}} = \vec{\mathbf{P}}^B$, the new prices.[7] $U^B > U^A$ iff

$$M\left(\vec{\mathbf{P}}^B; U^B\right) > M\left(\vec{\mathbf{P}}^B; U^A\right) \tag{9.21}$$

$$-T^B > M\left(\vec{\mathbf{P}}^B; U^A\right) \tag{9.22}$$

or

$$-T^B < -M\left(\vec{\mathbf{P}}^B; U^A\right) \tag{9.23}$$

But,

$$-T^A = M\left(\vec{\mathbf{P}}^A; U^A\right) \tag{9.24}$$

Thus, $U^B > U^A$ if

$$T^B - T^A < M\left(\vec{\mathbf{P}}^A; U^A\right) - M\left(\vec{\mathbf{P}}^B; U^A\right) \tag{9.25}$$

The left-hand side of Eqn (9.25) is the additional lump-sum subsidy required for the decreasing cost service to cover its full costs. It corresponds to distance $0b$ in Fig. 9.17. The right-hand side gives the income the consumer is willing to sacrifice to face prices $\vec{\mathbf{P}}^B$ instead of $\vec{\mathbf{P}}^A$. Thus, it corresponds to distance 0c in Fig. 9.17.

The right-hand side also represents a summation of areas under compensated demand (supply) curves. This interpretation follows from Shepard's lemma, that the partial derivative of the expenditure function with respect to the ith price is the compensated demand (supply) for good (factor) i. Therefore, letting all prices change one at a time:

$$M\left(\vec{\mathbf{P}}^A; U^A\right) - M\left(\vec{\mathbf{P}}^B; U^A\right) = \sum_{i=1}^{N} \int_{P_i^B}^{P_i^A} \frac{\partial M\left(\vec{\mathbf{P}}; U^A\right)}{\partial s} \mathrm{d}s$$

$$= \sum_{i=1}^{N} \int_{P_i^B}^{P_i^A} X_i^c \mathrm{d}P \tag{9.26}$$

When the ith compensated demand (supply) is integrated, it is evaluated at the prices P^B for the 1 to $(i-1)$ goods and factors that have already been integrated, and at prices P^A for the $(i+1)$ to N goods and factors that have yet to be integrated. Since the X_i are the compensated demands (supplies), the order of integration makes no difference.

If the new product is the first good and it is "small" so that prices P_i for $i \geq 2$ remain unchanged, then

$$M\left(\vec{\mathbf{P}}^A; U^A\right) - M\left(\vec{\mathbf{P}}^B; U^A\right) = \int_{P_i^B}^{P_i^A} X_i^c \mathrm{d}P_1 \tag{9.27}$$

where P_1^A is the price at which the demand curve intersects the price axis, and P_1^B is the marginal cost price. Hence, the area defined by Eqn (9.27) is EP_1^B A in Fig. 9.18, providing D refers to the compensated demand curve. Note, also, that the compensated demand curve lies to the left of the actual demand curve because the consumer sacrifices income as price is lowered to remain at the initial utility level. Thus the area behind the compensated demand curve corresponding to EP_1^B A in Fig. 9.18 is less than area EP_1^B A, in general.

Marshallian Consumer's Surplus and HCV

Needless to say, the hard case poses a number of practical difficulties for the policy maker. Operating the service at *any* single price generates losses, and justifying its continued operation in the face of these losses requires knowing a hypothetical willingness-to-pay income measure, such as HCV, that the general public will not understand. The public thinks in terms of profitability.

In addition, the HCV may not be easy to estimate even if it were understood. At best, the policy maker may know the aggregate market demand curve, although even this is extremely unlikely for many decreasing cost services. The all-or-none test requires knowing the demand relationship over the entire range of prices, from $P = \mathrm{MC}$ up to a price high enough to preclude any demand for the service. Yet, some of the hard-case decreasing cost services such as rural highways and some of the national parks have never been priced, so the quantities demanded at higher prices are unknown. In these cases, the econometrician is forced to use indirect methods to estimate the value of these services to consumers. When prices do exist, such as for mass rail transit, econometric analysis typically provides information on just a portion of the curve estimated over a relatively narrow range of historical prices. In some instances, there may be reasonable ways to extrapolate the estimated relationship back to the price axis, but even so one is left with the actual demand curve, not the compensated demand curve.

7. The expenditure functions could also be evaluated at $\vec{\mathbf{P}}^A$, which would involve Hicks' Equivalent Variation rather than Hicks' Compensating Variation.

Suppose enough price data exist to estimate the market demand curve along the full range of prices with a reasonable degree of confidence. Additional assumptions are still required to estimate the HCV. First and foremost is the assumption that income is continuously and optimally redistributed so that the first-best interpersonal equity conditions are satisfied and the economy is one-consumer equivalent. This assumption justifies the estimation of an aggregate market demand curve rather than individual demand curves because it assumes a well-defined set of social indifference curves defined over aggregate goods and services. Conversely, without this assumption the policy environment is second best and an appropriate second-best model would have to be specified and solved to determine the proper all-or-none second-best test. The literature offers a number of choices under the one-consumer equivalent assumption.

Jorgenson–Slesnick Expenditure Shaves

One possibility is the Jorgenson–Slesnick approach to demand estimation discussed in Chapter 4, in which expenditure shares are estimated in such a way as to recover the underlying (aggregate) indirect utility function in prices and income. Once the indirect utility function is obtained, it can be used to calculate the income required to hold utility constant as prices vary. The Jorgenson–Slesnick approach requires estimating an entire demand system, however, for which the data may not be available.

Roy's Identity

Another approach is to estimate a single market demand curve for the decreasing cost service and then make use of Roy's identity to recover the corresponding indirect utility function. Roy's identity states

$$\partial V(P,Y)/\partial P_i = \lambda X_i(P,Y) = \partial V(P,Y)/\partial Y \cdot X_i \quad (9.28)$$

where V is the indirect utility function, and X_i is the actual demand curve. Equation (9.28) is a differential equation in P_i and Y, which has a closed-form solution for V for certain demand functions. For example, the linear demand curve, $X_i = \alpha + \beta P_i + \gamma Y$ yields the indirect utility function:

$$V(P,Y) = e^{\lambda P}\left(\alpha/ - \beta/\gamma^2 + \beta/\gamma P + Y\right) \quad (9.29)$$

This example assumes that only the price of the service, P_i, is changing.[8] The remaining prices are part of the constant term a. Unfortunately, large projects such as mass rail transit systems or highways are likely to cause many prices to change. The single price assumption is highly suspected for the hard-case decreasing cost services.

8. We first saw this approach applied in Hausman (1981, fn. 19).

Marshallian Consumer Surplus

Still another popular approach is to assume away income effects so that the actual and compensated demand curves are one and the same. In this case, Marshallian consumer surplus and the appropriate willingness-to-pay income measures such as the HCV are identical, so there is no need to uncover the underlying indirect utility function. Assuming away income effects is hardly an attractive assumption, however. Almost all goods have some income elasticity of demand, and for services such as highways and recreational facilities, it may well be substantial. The higher the income elasticity of demand, the more these two benefit measures will diverge.

Nonetheless, Marshallian consumer surplus has remained a popular measure of the value of price changes, thanks to an approximation formula due to Robert Willig. Willig demonstrated that Marshallian consumer surplus is likely to be a close approximation of the HCV, even for fairly large income elasticities. Specifically, he proved that, for a single price change (Willig, 1976):

$$\frac{C-A}{A} \approx \frac{\eta A}{2M^0} \quad (9.30)$$

where

C = HCV due to the price change.
A = Marshallian consumer surplus.
η = income elasticity of demand.
M^0 = income in the original, no-service situation.

As Willig points out, if the surplus (A) is 5% of total income (M^0), even with an income elasticity (η) as high as 0.8, the error in using A for C is approximately 2%, well within the range of demand estimation error.

Willig's approximation formula is not without its problems. The assumption of a single price change is crucial to Willig's proof. If more than one price changes so that Eqn (9.26) applies, the Marshallian measure is not path dependent and is therefore not well defined. As noted above, the single-price-change assumption is highly suspect.

Roy's Identity Again

Peter Hammond has argued strongly, and persuasively, that applied economists should reject Marshallian consumer surplus measures of willingness-to-pay, Willig's approximation formula notwithstanding. Suppose an estimated demand curve does not lead to a closed-form solution for the indirect utility function or one of the willingness-to-pay income measures. Even so, Roy's identity can be used to construct an ordinary differential equation in prices and income from any estimated actual demand curve. The equation can then easily be solved by today's computers using standard numerical methods. Hammond demonstrates

the technique in terms of a solution that yields Hicks' Equivalent Variation (HEV), which is a valid income measure of the change in utility. Furthermore, the technique can be applied to any number of price changes. Modern computing has simply rendered Marshallian consumer surplus obsolete.[9]

Note, finally, that the problem of estimating the HCV or HEV to justify a decreasing cost service arises only for the hard case. Simply knowing that the service could break even or make a profit at a single price is sufficient in the easy case. As such, the profitability test that the public is familiar with applies even if the service is priced at marginal cost and operated at a deficit.

Decreasing Cost Services and Public Goods

A brief discussion of the relationship between decreasing cost services and nonexclusive public goods would be useful at this point because there is some confusion between the two in the professional literature. Decreasing cost goods with zero (or approximately zero) marginal costs are sometimes referred to as "public goods" because consumption is nonrival: Any one person's consumption of the good does not diminish the quantity available for others to consume.[10] The uncongested highway, bridge, or tunnel; national wildlife preserves; television viewing; and downloaded software are all reasonably good examples. Yet, referring to these services as "public goods" because marginal costs are (approximately) zero is extremely misleading. Samuelson coined the phrase "public good" for a particular kind of externality-generating activity, the nonexclusive good, for which consumption is nonrival because everyone necessarily consumes its services in equal amounts. In Chapters 6 and 7, we suggested an alternative definition of a public good that could also be applied to exclusive goods, providing the externalities generated by their consumption (production) affect everyone. Global warming is probably an example.

People are free to call things what they wish, but we believe the term "public good" ought to be reserved for certain kinds of externalities and not brought into the realm of decreasing cost theory. To apply it as well to decreasing costs, even when marginal cost is zero and consumption is nonrival, is bound to cause confusion. The problem is that the pareto-optimal rules for externality-generating activities differ substantially from their decreasing cost counterparts.

Consumption externalities lead to pareto-optimal rules of the form $\Sigma\ \text{MRS} = \text{MRT}$. Decreasing cost services, in contrast, requires the normal competitive rules, $\text{MRS} = \text{MRT}$, for pareto optimality. Furthermore, the marginal production costs of externality-type "public goods" need not be zero. The marginal costs for weapons systems are obviously considerable. And even if the marginal cost of a nonexclusive good happened to be zero, it would not be allocated the same way as a decreasing cost good with zero marginal cost.

To see that these two rules imply two distinct allocation mechanisms, compare the pareto-optimal allocations of an externality-type "public good" and a decreasing cost service in a two-person economy given that:

1. The marginal costs of providing each good are zero at every output.
2. The individual demand curves for each good are identical (but the two people have different demand curves).

These relationships are shown in Fig. 9.19.

If marginal costs (MRT) are zero for the decreasing cost good, then both persons should be allowed to consume the good until their personal MRSs are zero. The aggregate demand curve D_{DC}^{Total} is the *horizontal* summation of d_1 and d_2, and the optimum quantity is X_{DC}^{opt}, the point at which D_{DC}^{Total} intersects the X-axis. If marginal costs (MRT) are zero for the externality-type "public good," then the proper allocation occurs when $\text{MRS}^1 + \text{MRS}^2 = 0$. The aggregate demand curve D_{PG}^{Total} is the *vertical* summation of d_1 and d_2, and the optimum quantity is X_{PG}^{opt}, the point at which D_{PG}^{Topt} intersects the X-axis.

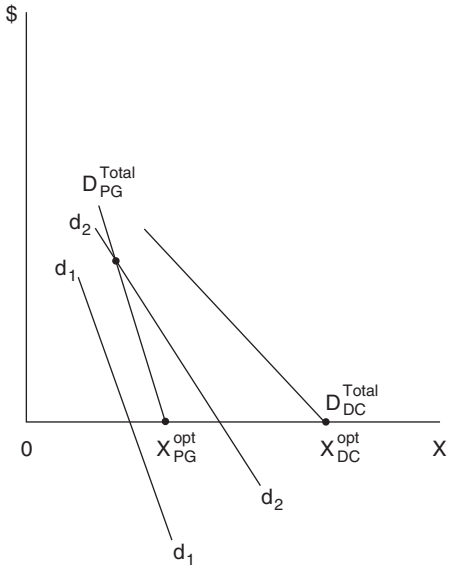

FIGURE 9.19

9. Hammond (1990), sections 10–12. Hammond also discusses aggregation problems when the distribution of income is not optimal so that the economy is not one-consumer equivalent.

10. Francis Bator describes "publicness" in this manner in *The Question of Government Spending:* "There are activities, however, where the additional cost of extra use is literally zero. The economist labels the output of such activities 'public'." See Bator (1960, p. 94). (See pp. 90–98 for an extended discussion of zero marginal cost and decreasing cost services, especially pp. 94 and 96.)

Finally, the decreasing cost good can be marketed more or less normally since each consumer should face the same price (=0). Of course, the government does have to ensure that the fixed costs are covered with a lump-sum subsidy, but this is true of any properly marketed decreasing cost service, even those with nonzero marginal costs. In contrast, "marketing" the nonexclusive public good requires that the government selects the single quantity.

The consumers may then be charged their demand prices (Lindahl prices) at the chosen quantity, but this is not necessary. Any lump-sum tax preserves pareto optimality.

In conclusion, the nonrivalry quality that "any one person's consumption does not diminish the quantity available to anyone else" is not precise enough to be useful. It could refer to a nonexclusive good or it could imply nothing more than zero marginal costs (MRT). To avoid this ambiguity, we believe that the term "public good" should be reserved for instances of pervasive externalities, more or less as Samuelson originally intended. If the term is also used to characterize zero marginal cost, decreasing cost services, it loses its particular analytical significance. It might as well refer to any good requiring government intervention, since there is no analytical reason to distinguish between zero and nonzero marginal cost decreasing cost services.

REFLECTIONS ON U.S. POLICY REGARDING DECREASING COST SERVICES: THE PUBLIC INTEREST IN EQUITY AND EFFICIENCY

Suppose a decreasing cost service satisfies the requirements of the "easy case," that a profit-maximizing monopolist could at least break even. The easy case presents three obvious pricing options for the government, each depicted in Fig. 9.20.

The simplest option is to preserve free enterprise, offer the natural monopoly to private investors, and let them operate the service as a monopolist. The expectation is that the owners will choose to maximize profits by producing output X_m at which MR = MC, setting price equal to P_m, and earning pure profits of $(P_m - \text{AC}) \cdot X_m$.

The other two options involve government intervention, either in the form of a direct government takeover of the service or private ownership with government regulation. In either case, the government has two natural choices:

1. Follow the dictates of first-best theory, charge the marginal cost price P_{MC}, and subsidize the operation out of general tax revenues in the amount of $(\text{AC} - \text{MC}) \cdot X_{\text{opt}}$.
2. Charge a price equal to average costs, P_{AC}, and produce X_{AC}, in which case the service covers its full costs.

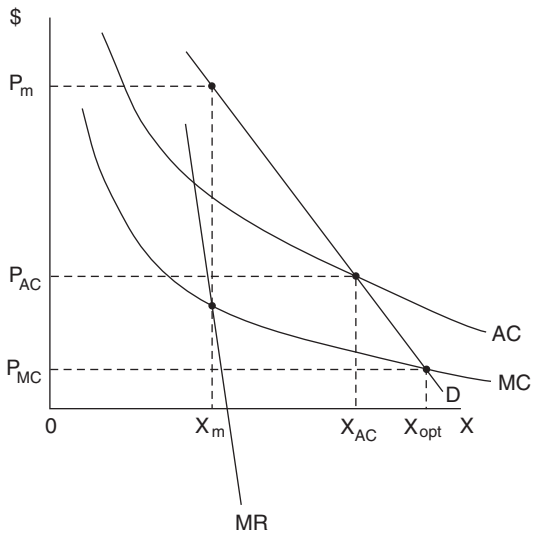

FIGURE 9.20

U.S. governments at all levels have overwhelmingly adopted the average cost pricing strategy, or some close approximation to it, whether the service is publicly or privately owned. For example, fees for recreational facilities such as beaches and parks are usually set to cover the full costs of operating these facilities. Tolls on urban highways, bridges, and tunnels are often designed to cover the full costs of the entire network of transportation facilities under the jurisdiction of a local transportation authority. The federal gasoline tax was originally established to defray the expenses of constructing the interstate highway system. Similarly, state gasoline tax rates are determined primarily by the anticipated expenses of state highway departments. Public utility rates are generally designed to cover all expenses including a fair rate of return to the private investors. Admittedly, unless the "fair return" equals the opportunity cost of capital services, this is not strictly average cost pricing, but its philosophy is more or less identical. One can think of the utility-regulatory commissions as constructing an average cost curve that includes the "fair" rate of return and setting rates equal to these constructed average costs.

In some instances, governments have not insisted on average cost pricing for decreasing cost services. Examples include some national and state parks and beaches financed out of general revenues, over-the-air commercial television financed by advertising revenues, and sales of rights to use recorded music through agencies such as ASCAP and BMI, which charge users a onetime annual fee for access to their music inventories. Notice that in each of these examples, the per-use price of the service is essentially zero. Since the marginal cost of these services is likely to be near zero, where the quantity axis defines the number of users or viewers, the zero price can be thought of as roughly

consistent with optimal pricing (so long as the service remains uncongested). The use of general revenues, advertising, or onetime fees to cover costs may not be optimal, however.

The preference for average cost pricing may seem surprising, given the general support for free enterprise in the United States or the first-best theoretical arguments favoring marginal cost pricing with government intervention. The question arises as to why average cost pricing is so common in the United States. What are its perceived advantages and to what extent are these perceptions reasonable?

In our view, the United States accepts the average cost option as a reasonable compromise between the other policy options on both equity and efficiency grounds. We would also hazard a guess that equity issues are the more compelling to the general public, but efficiency arguments are at least considered in most public deliberations on price setting.

Equity Considerations

The interesting equity issue concerns the choice between average cost pricing and marginal cost pricing. U.S. citizens will not willingly permit a private owner to "exploit" a natural monopoly position and earn monopoly profits. Dissatisfaction in the United States over public price increases that are ostensibly justified by cost increases is often severe. One can well imagine the public's outrage over a charge of profiteering at the public's expense.

Perhaps, the outstanding example of the government hedging against profiteering occurs in defense contracting.[11] Complex weapon systems routinely experience huge cost overruns. One of the more obvious reasons why is that the government negotiates cost-plus-fixed-fee contracts through the research and development stages of the production cycle, so that there is little incentive for contractors to hold down cost. An equally obvious solution is to insist on fixed-fee contracts from the beginning, but given initial uncertainties over cost and quality parameters the government has been willing to use them only sparingly. Apparently, the federal government considers the risk of huge profits (and huge losses) for its few large weapons suppliers less acceptable than the cost overruns, despite incessant public disfavor with the latter.[12]

Although the defense contractors are not decreasing cost industries, the same principle undoubtedly applies to the decreasing cost services as well. The fear of the private owners profiteering at the public's expense is probably sufficient to rule out the private monopoly option. One might argue that a natural monopoly would not fully exploit its monopoly power knowing that excessive profits would not be tolerated. For example, although ticket prices are usually raised for important sporting events (e.g., baseball's World Series, football's Super Bowl), public pressure clearly keeps owners from setting even higher, market-clearing prices. In any event, average cost pricing avoids profiteering, at least in principle. As a practical matter, it is questionable whether a monopoly such as a public utility can be effectively regulated to avoid all monopoly profit.

The more subtle question is why governments favor average cost over marginal cost pricing, despite the obvious efficiency advantages of the latter. We believe that the answer lies in the public's belief in the benefits-received principle of taxation or public pricing, a principle that was first discussed in Chapter 6 in the context of paying for nonexclusive goods.

Recall that the benefits-received principle is commonly accepted as an equity principle,[13] which, broadly interpreted, states that consumers should pay for a public service in direct proportion to the benefits they receive. A natural corollary is that those who receive no benefits should not have to pay for the service. Suppose the government chooses the marginal cost pricing option for a decreasing cost service that it operates or regulates. The marginal cost price itself is consistent with the benefits-received principle: Only users pay the price and more intensive users pay it more often. The problem comes with the subsidy required to cover the losses, which is presumably paid out of general tax revenues. Consumers no longer pay for the full costs of the service in proportion to their use of the service when a substantial portion of the costs are covered by general tax revenues. Contributions to the subsidy are more likely to be proportional to income or consumption than to the use of the service. Worse yet, some nonusers may end up paying part of the costs with their taxes.

As indicated in Chapter 6, the benefits-received principle begs the issue of which benefits the payments ought to be related to. Even so, the easy-case decreasing cost services are perfect candidates for pricing according to benefits received. They are exclusive goods for which nonusers can easily be distinguished from users and more-intensive users from less-intensive users. Moreover, a single price can cover the average cost of the service. Therefore, one could reasonably argue that average cost pricing satisfies the intent of the

11. The classic references on defense contracting are Peck and Scherer (1962) and Scherer (1964).

12. The same issues are being revisited in the debate over the best way to provide health care. The fee-for-service payment for hospitals and physicians has undoubtedly contributed to the steady rise in medical costs. The HMO single-payment alternative does better at containing costs, but critics complain that the HMOs cut corners on care to increase the profits of their investors.

13. Recall also from the discussion of Chapter 6 that the benefits-received principle is not a valid equity principle within the formal neoclassical model, despite its long standing within the profession. All end-results equity considerations are incorporated into the interpersonal equity conditions for a social welfare maximum.

benefits-received principle, whereas marginal cost pricing does not. If people adopt this point of view and believe that the equity gains from average cost pricing outweigh its efficiency losses relative to marginal cost pricing, then average cost pricing is entirely reasonable.

Economists can easily attack this position by appealing to first-best theoretical principles, but it is not at all clear that the general public will find the economic case very compelling. Recall that the economic argument runs as follows. The benefits-received principle is meant to be applied to all exhaustive or resource-using government expenditures, those undertaken to correct for misallocations of resources within a competitive market system. The public may view it as an equity principle in the sense that it imitates the *quid pro quo* payment mechanism of the free-market system, but its real purpose is to support an efficient allocation of resources, exactly as competitive prices do in markets for which all the technical and market assumptions hold.

For nonexclusive goods, an infinity of payment schemes preserves pareto optimality, but not for decreasing cost services. Only if the benefits-received principle is interpreted to mean marginal cost pricing is its efficiency function upheld. Marginal cost pricing is, of course, also consistent with the equity criterion that it should imitate competitive pricing. Each person is allowed to consume the good until price equals MRS (assume that the other good is the numeraire), which in turn is equated to the MRT in production. According to this interpretation, then, payment is related to use only to the point at which price covers marginal cost.

Although the benefits-received principle so interpreted does not cover the full costs of decreasing cost services, this is simply irrelevant. According to first-best theory, payment of the required subsidy through lump-sum taxes depends only on the interpersonal equity conditions of social welfare maximization. It has nothing to do with use or nonuse of the service. Those people who ultimately support these services through lump-sum taxes are simply those who originally have relatively low social marginal utilities of income (e.g., the rich). Conversely, the people whose use of the service is implicitly subsidized by the set of lump-sum taxes and transfers receive implicit subsidies only because they have relatively high social marginal utilities of income (e.g., the poor).

To clarify this point, suppose a decreasing cost service is paid for by an efficient two-part tariff consisting of a marginal cost price for actual use and a onetime, lump-sum fee collected from all actual and potential users of the service.[14] This onetime fee may seem desirable from a benefits-received perspective because the users pay the full costs.[15] But the government's distribution bureau will effectively override the lump-sum payments if they do not square with the interpersonal equity conditions.

Suppose, for example, that only poor people use rail transit and that the interpersonal equity conditions require a net redistribution from the rich to the poor. The lump-sum user fees drive the social marginal utilities of income of the rich and the poor further apart. The distribution bureau simply offsets this, however, by taking even more income from the rich and transferring it to the poor until their social marginal utilities are equalized. Although the poor users appear to be covering the transit costs, the rich actually are, precisely because they are rich. The interpersonal equity conditions always take precedence in social welfare maximization.

The marginal cost interpretation of the benefits-received principle may be consistent with first-best principles, but the general public is not likely to accept it, especially its equity implications. Subsidizing a public service out of general revenues would undoubtedly be highly unpopular, even though it would permit lower prices for using the service. The benefits-received principle is deeply ingrained as an equity principle in the United States.

In summary, we are quite prepared to admit that an average cost interpretation of the benefits-received principle squares best with the public's notion of equity in the provision of decreasing cost services. First-best theory notwithstanding, economists should perhaps concede that average cost pricing has certain appealing equity properties.

Efficiency Considerations

The efficiency advantages of marginal cost pricing over average cost pricing are unambiguous in a first-best environment, since the marginal cost price is pareto optimal. Nonetheless an "easy-case" service at least passes an all-or-none efficiency test if priced at average cost. We saw that operating the service at the break-even output is preferable to having no service at all. When the average cost pricing philosophy is applied to a "hard-case" service, however, society risks having the service fail even this gross efficiency test.

Strict average cost pricing is impossible in the hard case, of course, because demand is everywhere below average cost. In lieu of covering the full costs, the public may insist on minimizing the deficit as the next-best alternative: If the users cannot cover the full costs of the service, they at least should pay for as much of the costs as possible.

14. The onetime fee must be collected from potential users, or an economic choice to use or not use the service would dictate the amount of payment, contrary to the notion of a lump-sum payment. We saw this same problem when considering subsidies for nonpolluting behavior in Chapter 7.

15. Potential but not actual users can be thought of as purchasing an option to use the service, which, if they pay the fee, must have value to them.

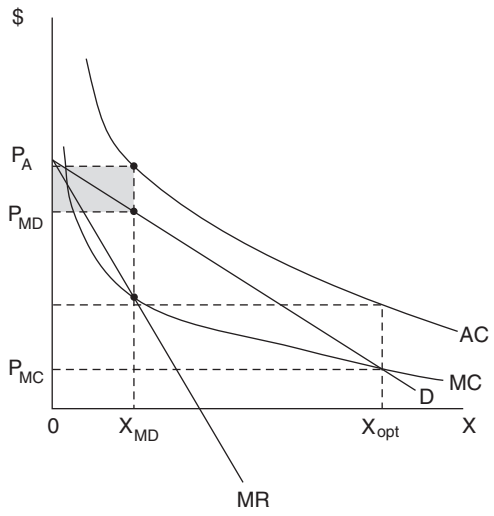

FIGURE 9.21

Unfortunately, minimizing the deficit may not be a harmless extension of the average cost principle. The minimum deficit solution may be grossly inefficient.

Refer to Fig. 9.21. Minimizing the deficit is the same as maximizing profits in the hard case. Therefore, the level of service that minimizes the deficit is X_{MD}, at the intersection of the MR and MC curve, with a price of P_{MD}. The shaded area is the minimum deficit.

Suppose the service passes the all-or-none test if priced at marginal costs and operated at X_{opt}. Even so, it could fail the all-or-none necessary condition at (P_{MD}, X_{MD}). That is, it is possible for HCV from P_A to P_{MC} to exceed the lump-sum subsidy required to cover the deficit at X_{opt}, whereas the HCV from P_A to P_{MD} fails to cover the deficit at X_{MD}.[16] This is certainly the case in Fig. 9.21. The mere potential for satisfying the all-or-none test at the optimum is not enough. Unless society actually operates the service at a level sufficient to pass this test, it is simply wasting resources. Minimizing the deficit is not the same principle as maximizing the difference between total benefit and total cost, because total benefit equals total revenue plus the HCV.

Rail transit in a number of urban areas could be an example of the dangers of the minimum deficit philosophy, although we do not know enough about either transit demand or costs to say for sure. Despite numerous fare increases designed specifically to eliminate deficits (presumably demand is inelastic in the relevant range), the deficits persist and, predictably, ridership diminishes. We speculated earlier that rail transit may be an example of the "hard case" such that no single fare can avoid an operating deficit. In addition, if ridership continues to decline as fares are increased to lower the deficit, the transit system may not be worth operating at all. The trains may run too empty too often.

One can imagine the public outcry at a suggestion to lower fares and incur even larger deficits, even if this were the only way that the transit service could pass the all-or-none test. The general public is unlikely to understand the subtleties associated with the hard-case decreasing cost services. Their belief in profitability as the proper guide for the use of scarce resources is deeply held, and understandably so. They see the profitability test applied every day in the marketplace.

16. For the purposes of this discussion, assume zero income effects, so that $D^{actual} = D^{compensated}$.

APPENDIX: RETURNS TO SCALE, HOMOGENEITY, AND DECREASING COST

Since *increasing returns to scale* implies *decreasing average cost*, the two terms are used interchangeably in the chapter. To see that the former implies the latter, consider the homogeneous production function:

$$Y = f(X_1, \ldots, X_N) = f(X_i) \tag{9A.1}$$

where X_i = input i, $i = 1,\ldots, N$, and Y = output. By the definition of homogeneous functions,

$$\lambda^B Y = f(\lambda \cdot X_i) \tag{9A.2}$$

Increasing returns to scale implies that $\beta > 1$, or a scalar increase (decrease) in each of the factors generates a more-than-proportionate increase (decrease) in output. Furthermore, if the production function is homogeneous of degree β, then the marginal product functions, $\partial Y/\partial X_k \equiv f_K(X_i)$ are homogeneous of degree $\beta - 1$. This follows immediately by differentiating $\lambda^\beta Y = \lambda^\beta f(X_i) = f(\lambda X_i)$ with respect to X_k:

$$\lambda^B f_k(X_i) = \frac{\partial f(\lambda X_i)}{\partial X_K} = \lambda f_K(\lambda X_i) \tag{9A.3}$$

Hence,

$$\lambda^{\beta-1} f_K = f_K(\lambda X_i) \quad k = 1, \ldots, N \tag{9A.4}$$

To minimize production costs for any given output, the firm solves the following problem:

$$\min_{(X_i)} \sum P_i X_i$$
$$\text{s.t. } Y = f(X_i)$$

The first-order conditions imply

$$\frac{P_i}{P_l} = \frac{f_i(X_i)}{f_l(X_i)} \quad i = 2, \ldots, N \tag{9A.5}$$

The ratio of factor prices equals the marginal rate of technical substitution of the factors in production. Suppose the firm increases (decreases) its use of all factors

X_i by the scalar λ. Since $f_i(\lambda X_i) = \lambda^{\beta-1} f_i(X_i)$, this scalar increase (decrease) continues to satisfy the first-order conditions:

$$\frac{f_i(\lambda X_i)}{f_i(\lambda X_i)} = \frac{\lambda^{\beta-1} f_i(X_i)}{\lambda^{\beta-1} f_1(X_i)} = \frac{f_i(X_i)}{f_i(X_i)} = \frac{P_i}{P_1} \tag{9A.6}$$

Hence, if a vector of inputs $\vec{\mathbf{P}}_{*i}$ minimizes cost, so too will any vector $\lambda \vec{\mathbf{P}}_{*i}$. But, if all inputs are increased by the scalar λ, total costs increase by λ and output increases by a factor λ^{β}. Thus, the total cost function must be of the form:

$$\mathrm{TC} = kY^{1/\beta} \tag{9A.7}$$

since $k \cdot \left(\lambda^{\beta} Y\right)^{1/\beta} = \lambda \cdot k \cdot Y^{1/\beta} = \lambda \cdot \mathrm{TC}$. Finally,

$$\mathrm{AC} = \mathrm{TC}/\mathrm{Y} = kY^{(1/\beta-1)} = kY^{(1-\beta)/\beta} \tag{9A.8}$$

Differentiating,

$$\frac{\partial \mathrm{AC}}{\partial \mathrm{Y}} = \frac{1-\beta}{\beta} k \cdot Y^{((1-\beta)/\beta-1)} < 0, \text{ for } \beta > 1 \tag{9A.9}$$

Hence, average cost declines continuously as output increases with increasing returns to scale.

REFERENCES

American Public Transportation Association, August 02, 2004. Table 135, www.apta.com.

Bator, F.M., 1960. The Question of Government Spending. Harper Brothers, New York.

Diamond, P., McFadden, D., February 1974. Some uses of the expenditure function in public finance. Journal of Public Economics 3 (1), 3–21.

Hammond, P., January 1990. Theoretical progress in public economics: a provocative assessment. Oxford Economic Papers 42 (1), 6–33.

Hausman, J., 1981. The effect of taxes on labor supply. published as a chapter entitled "Labor Supply". In: Aaron, H., Pechman, J. (Eds.), How Taxes Affect Economic Behavior. Brookings, Washington, D.C.

Peck, M., Scherer, F., 1962. The Weapons Acquisition Process: An Economic Analysis. Division of Research, Graduate School of Business Administration, Harvard University, Boston, MA.

Scherer, F., 1964. The Weapons Acquisition Process: Economic Incentives. Division of Research, Graduate School of Business Administration, Harvard University, Boston, MA.

Willig, R., September 1976. Consumer's surplus without apology. American Economic Review 66 (4), 589–597.

Chapter 10

The First-Best Theory of Taxation and Transfers

Chapter Outline

Having covered the mainstream normative theory of public expenditures in Chapter 2 through 9, the mainstream first-best theory of taxation is easy to describe. We saw that first-best public expenditure theory addresses two fundamental questions:

1. In what area of economic activity can the government legitimately become involved?
2. What decision rules should the government follow in each area?

In answering these questions, public expenditure theory provides both a complete prescription for government expenditures and a complete normative theory of taxation. There is no first-best theory of taxation distinct from the first-best theory of public expenditures. All we need do is review the main results from the previous chapters.

The first point to recall is that taxes can only enhance social welfare in a first-best environment. They are not at all the necessary evil that the public sometimes makes them out to be. On the contrary, first-best taxes promote the public interest in efficiency and equity as they support society's quest for a social welfare maximum at the bliss point. They have no other purpose in the mainstream first-best theory.

Regarding efficiency, public expenditure theory either describes some particular tax necessary to achieve a pareto-optimal allocation of resources or it does not. If not, then taxes have no further role to play in promoting economic efficiency. For example, we found that exclusive goods that generate either consumption or production externalities can be allocated correctly with a set of Pigovian taxes (subsidies) equal to the aggregate marginal damage (benefit) resulting from the externality. Similarly, decreasing-cost services require marginal cost pricing for pareto optimality. Whether one refers to these publicly set prices as admission "fees," highway and bridge "tolls," or transit "fares" hardly matters. The marginal cost charges for these services can always be thought of as taxes set according to the competitive interpretation of the benefits-received principle of taxation, the only interpretation consistent with economic efficiency.

One can analyze the efficiency costs of distorting taxation, of course. In fact, the analysis of distorting taxation dates from the very beginnings of modern public finance when taxes received far more attention than expenditures. But distorting taxation is inherently part of the second-best theory.

At times first-best public expenditure theory requires certain government expenditures without specifying exactly how to collect the revenues to finance these expenditures. Leading examples are Samuelsonian nonexclusive public goods and subsidies to cover the deficits of decreasing-cost services when prices (taxes) are set equal to marginal costs. The only efficiency criterion in these instances is that the taxes be *lump sum* to avoid generating distortions that would prevent the first-best pareto-optimal conditions from holding. Any pattern of lump-sum taxation preserves the efficient allocation of these goods.

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00010-8

As we have seen, first-best theory also solves the problem of how to collect the lump-sum taxes to finance these expenditures. The required taxes (transfers) simply become part of the pattern of lump-sum taxes and transfers that satisfy the interpersonal equity conditions of social welfare maximization of the form

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial I^h} = \text{all} \quad h = 1, \ldots, H$$

where I^h can loosely be thought of as lump-sum income.[1] Whether the requirement of equalizing the social marginal utilities of income is viewed as part of first-best expenditure theory or first-best tax theory is a matter of semantics since lump-sum taxes and transfers are analytically equivalent except for the sign. Either way, the interpersonal equity conditions are the only equity criterion for taxes and transfers in a first-best policy environment. They represent, simultaneously, a complete prescription for the optimal distribution of income and for the optimal redistribution of income when starting from a nonoptimal distribution.

Having considered how taxes and transfers help achieve the pareto-optimal and interpersonal equity conditions of first-best social welfare maximization, mainstream public sector theory has nothing more to say about them in a first-best environment.

PUBLIC CHOICE AND PARETO-OPTIMAL REDISTRIBUTION

The policy implications associated with the interpersonal equity conditions are perhaps the least convincing component of the mainstream first-best theory. On the one hand, any tax or transfer that is related to an individual's well-being, such as a personal income tax or a means-tested transfer payment, is unlikely to be lump sum. On the other hand, the interpersonal equity conditions rely on the social welfare function, which, although a useful analytical construct, is problematic in the extreme as a practical guide to distributional policies. As noted in Chapter 3, there is no convincing theory to determine what the marginal social welfare weights should be, and Arrow proved that a democratic society may not be able to articulate a consistent social welfare function when individuals disagree about the appropriate ethical marginal social welfare weights.

Mainstream public sector economists have responded to the difficulties of the social welfare function in one of two ways, for the most part. The first might be called the technocratic response, most closely associated with Paul Samuelson. The idea here is to concede that economic theory has nothing useful to say about the form of the social welfare function. At the same time, an operative social welfare function undoubtedly exists; the ruling politicians are setting their policies with some set of marginal social welfare weights in mind. Therefore, simply ask the policy makers what their social welfare function is and then tell them what the optimal policies are given that function. Economists can solve constrained optimum problems once they know what the objective function is. The second mainstream response, and the more common one, is to retain the social welfare function in normative analysis but use a flexible form of the function that permits a wide range of social welfare weights. The Atkinson and Jorgenson social welfare functions described in Chapter 4 are examples. They each employ Atkinson's aversion to inequality parameter, which admits the full range of social welfare weights that people are likely to prefer, from utilitarian indifference to Rawlsian egalitarianism. The idea behind the flexible-form approach is to see how optimal policies vary with the social welfare weights.

A third, and very different, response to the social welfare function comes from the public choice economists following James Buchanan. They are unconcerned about the difficulties surrounding the social welfare function because they reject it out of hand. They see it as an illegitimate construct based on the patently false assumption that people are self-interested in their economic affairs and yet other-interested in their political affairs as they think about an appropriate social welfare function for resolving the distribution question. In their view, people are just as self-interested in their political affairs; they simply do not think in terms of a social welfare function.

The public choice position raises an interesting question. The United States spends approximately $825 billion a year (Fiscal Year 2013) in public assistance to the poor through such programs as Supplemental Security Income (SSI), Temporary Assistance to Needy Families (TANF), Medicaid, Supplemental Nutrition Assistance Program (SNAP, the Food Stamp program), Housing Assistance, and the Earned Income Tax Credit (EITC). Can such a large a public assistance effort possibly be explained without something like an other-interested, politically determined social welfare function? Yes, say the public choice economists. Just think of public assistance as a natural extension of private charity that governments undertake because of certain limitations in private giving. Their view of public assistance is as follows.

Private donations to the poor indicate that people are not narrowly self-interested in their private lives. They do have altruistic impulses toward the poor and are willing to help them. Moreover the private donations meet the test of a policy-relevant, technological consumption externality. The donations result because the plight of the poor directly affects the nonpoor donors (enters their utility functions), and the voluntary donations occur outside the normal market

1. More precisely, it is a good or factor (presumably the numeraire) singled out for taxation and transfer.

channels. Therefore, the optimal pattern of donations is determined by the standard kind of pareto-optimal conditions that apply to consumer externalities, not by the interpersonal equity conditions of an illegitimate social welfare function. According to the public choice perspective, the quest for end-results equity is entirely subsumed within the quest for efficiency.

Harold Hochman and James Rodgers were the first to formalize the notion of redistributive taxes and transfers from the perspective of consumer externality. They referred to the optimal policy as a *pareto-optimal redistribution*, a label that has stuck in the literature (Hochman and Rodgers, 1969).

The public choice economists see a distinct advantage in viewing the optimal distribution as an efficiency rule. It presumes that distributional policy is a self-interested gain−gain policy rather than an other-interested, lose−gain policy, consistent with economic rationality. The taxes and transfers driven by the interpersonal equity conditions of social welfare maximization imply that those who are taxed are willing to lose, to sacrifice some of their own utility for the greater good of supporting the poor. pareto-optimal redistribution, in contrast, is a gain−gain proposition—the donors as well as the recipients gain. This may seem like an unimportant distinction when donors are altruistic, but there are important differences from a political perspective. A lose−gain policy runs into the difficulties of determining how to compare the losses of the losers with the gains of the gainers. It is also vulnerable to Arrow's Impossibility Theorem in a democracy if people vary in their willingness to sacrifice for the poor. A gain−gain redistribution, in contrast, would presumably receive unanimous consent in a democratic election.

Pareto Optimality and the Overall Distribution of Income

The view of redistribution as a consumer externality applies so long as any concerns about the distribution of income enter into people's utility functions, not necessarily just concerns about the poor. Therefore, let us begin with the more general model of pareto-optimal redistribution specified in terms of the overall distribution, as Hochman and Rodgers did.

To keep the analysis as simple as possible, imagine an exchange economy in which each individual, h, has an endowment of two goods (factors), $\overline{Y}_h$ and $\overline{Z}_h$. The total supply of Y and Z is assumed fixed, equal to $\overline{Y} = \sum_{h=1}^{H} \overline{Y}_h$ and $Z = \sum_{h=1}^{H} \overline{Z}_h$.[2]

2. There is no need to model production since we are only concerned with distribution rules. Were production included it would have no effect on the optimal distributional decision rules in a first-best economy.

Suppose each person's utility is a function of Y, Z, and the distribution of Y among all members of the society. That is,

$$U^h = U^h\left(Y_h, Z_h, X\right)$$

where $X = X(Y_h)$ represents an index of the distribution of Y among all H individuals. Assume there is no social welfare function, in keeping with the public choice perspective. Instead, society's goal is to achieve the pareto-optimal allocation of Y_h and Z_h given the total fixed endowments of $\overline{Y}$ and $\overline{Z}$.

The pareto-optimal conditions are derived by maximizing the utility of any one person, say person 1, subject to holding the utility of all other people constant and to the endowment constraints. Formally,

$$\begin{aligned} &\max_{(Y_1, Z_1, Y_h, Z_h)} U^1\left(Y_1, Z_1, X\right) \\ &s.t. \quad U^h\left(Y_h, Z_h, X\right) = \overline{U}^h \quad h = 2, \ldots, H \\ &\sum_{h=1}^{H} Y^h = \overline{Y} \\ &\sum_{h=1}^{H} Z^h = \overline{Z} \end{aligned}$$

The corresponding Lagrangian is

$$\begin{aligned} \max_{(Y_1,\ Z_1, Y_h, Z_h)} L = {} & U^1\left(Y_1, Z_1, X\right) + \sum_{h=2}^{H} \lambda^h\left(U^h\left(Y_h, Z_h, X\right)\right) \\ & + \pi\left(\overline{Y} - \sum_{h=1}^{H} Y^h\right) + \delta\left(\overline{Z} - \sum_{h=1}^{H} Z^h\right) \end{aligned}$$

The first-order conditions are

$$Y_1 : \frac{\partial U^1}{\partial Y_1} + \frac{\partial U^1}{\partial X}\frac{\partial X}{\partial Y_1} + \sum_{h=2}^{H} \lambda^h \frac{\partial U^h}{\partial X}\frac{\partial X}{\partial Y_1} = \pi \qquad (10.1)$$

$$Y_i : \frac{\partial U^1}{\partial X}\frac{\partial X}{\partial Y_i} + \lambda^i \frac{\partial U^i}{\partial Y_i} + \sum_{h=2}^{H} \lambda^h \frac{\partial U^h}{\partial X}\frac{\partial X}{\partial Y_i} = \pi \quad i = 2, \ldots, H \qquad (10.2)$$

$$Z_1, Z_h : \frac{\partial U^1}{\partial Z_1} = \delta = \lambda^h \frac{\partial U^h}{\partial Z_h} \quad h = 2, \ldots, H \qquad (10.3)$$

Dividing Eqn (10.1) or (10.2) by Eqn (10.3), with appropriate selection of h in Eqn (10.3) yields:

$$\frac{\partial U^i}{\partial Y_i} \Big/ \frac{\partial U^i}{\partial Z_i} + \sum_{h=1}^{H} \left(\frac{\partial U^h}{\partial X}\frac{\partial X}{\partial Y_i} \Big/ \frac{\partial U^h}{\partial Z_h}\right) = \pi/\delta \quad i = 1, \ldots, H \qquad (10.4)$$

Equation (10.4) has the standard form for a consumption externality. It says that the government should equate each person's personal-use marginal rate of substitution

between Z and Y, plus the sum of everyone's marginal rate of substitution between their own consumption of Z and the effect of Y_i on the distributional index X. These rules are identical to the pareto-optimal rules for allocating exclusive pure public goods. The only difference is in their interpretation. They are distribution rules, the optimal policy for redistributing Y across the population. In other words, they are the recipe for a pareto-optimal redistribution.

Note, too, that because the optimal distribution rule is described in terms of marginal rates of substitution (MRSs), it can be achieved by competitive markets for Y and Z buttressed by a set of H personalized Pigovian taxes or subsidies on good (factor) Y:[3]

$$t_i = \sum_{h=1}^{H} \left(\frac{\partial U^h}{\partial X} \frac{\partial X}{\partial Y_i} \Big/ \frac{\partial U^h}{\partial Z_h} \right) \quad i = 1, \ldots H \qquad (10.5)$$

The taxes (subsidies) equal the aggregate marginal external effect of an additional unit of consumption of Y by person i, the standard interpretation of a Pigovian tax, with the external effect arising through the concern for the distribution.

The taxes and subsidies would be difficult to implement because the distribution is an example of an individualized externality. In principle, H taxes are required to achieve the pareto-optimal conditions. The policy burden would be lessened, however, if society thought of the distribution in terms of, say, deciles of the population and assumed that everyone within a given decile had the same effect on X.

Pareto-Optimal Redistribution and the Poor

The United States is unlikely to try to implement rules such as Eqn (10.4) because it has never articulated a policy regarding the overall distribution of income. There has always been a concern for helping the poor, however, which reached its zenith in 1964 when President Johnson declared a War on Poverty. The goal of the war effort was nothing less than the eradication of poverty in the United States, a goal that remains elusive. As of 2013, over 46 million people in the United States live in poverty.

The implications of pareto-optimal redistributions on antipoverty policies are best seen with a simpler version of the general distribution model above. Suppose that society consists of two classes of people, the rich (nonpoor) and the poor. The rich are concerned about the economic state of the poor, but are unconcerned about the distribution generally. The poor care only about their own consumption and utility; they have no concerns about distributional matters. A model of this form captures the motivation for private charity toward the poor.

Begin with the simplest possible two-person, two-good endowment model, consisting of one rich person, one poor person, and the two goods Y and F. Y is a composite commodity and F is food. The poor person's utility is a function of his own consumption of Y and F:

$$U^p = U^p(Y_p, F_p)$$

The rich person's utility is a function of her own consumption of Y and F and the poor person's consumption of food:

$$U^r = U^r(Y_r, F_r, F_p)$$

That is, when the rich person considers the plight of the poor person, her concern is that the poor person has enough to eat.

The first-order, pareto-optimal conditions for this simple model are easily shown to be

$$\text{MRS}^r_{Y_r, F_r} = \text{MRS}^p_{Y_p, F_p} + \text{MRS}^r_{Y_r, F_p} \qquad (10.6)$$

When the rich person consumes Y and F, only her personal-use MRS matters, the left-hand side of Eqn (10.6). Her consumption does not generate an externality. When the poor person consumes Y and F, however, two MRSs come into play: his personal-use MRS, the first term on the right-hand side (RHS) of Eqn (10.6), and the rich person's MRS between the poor person's consumption of food and her own consumption of Y, the second term on the RHS of Eqn (10.6). The second term indicates the amount of Y the rich person would be willing to sacrifice for the poor person to consume one more unit of food. The sum of the two terms on the RHS is the full social MRS of the poor person's consumption of Y and F, which must equal the rich person's personal-use MRS for a pareto optimum. The rich person would presumably transfer food to the poor person to achieve the pareto optimum in this simple world. Private charity would suffice without the need for government intervention.

The problems with private charity motivated by altruism arise because there are many rich people who care about the poor person. To see this, expand the model to include many rich people, each with the same utility function defined above. The pareto-optimal conditions for the expanded model are

$$\text{MRS}^r_{Y_r, F_r} = \text{MRS}^p_{Y_p, F_p} + \sum_r \text{MRS}^r_{Y_r, F_p} \quad \text{all } r \, \varepsilon \, R \qquad (10.7)$$

where R is the set of rich people. The second term on the RHS of Eqn (10.7) is the aggregate marginal external effect on the rich of the poor person's consumption of food.

3. $\partial X/\partial Y_i$ would be positive for some people (i.e., more equalizing) and negative for others (i.e., less equalizing). Z is the numeraire.

Condition (10.7) runs afoul of the free-rider problem when altruistic people are otherwise self-serving. Each rich person receives a boost in utility when the poor person consumes another unit of food, regardless of who supplies the food. Therefore, every rich person has an incentive to free ride: let someone else supply the food and thereby capture the utility gain at no cost. Alternatively, no rich person wants to play the sucker and provide the food for the benefit of all the other rich people.

The incentive to free ride drives charity into the public sector in the form of public assistance. The government can achieve the pareto optimum by following the standard Pigovian subsidy prescription for beneficial consumption externalities. Suppose the markets for Y and F are competitive, in line with first-best assumptions, with competitive prices P_F and P_Y, and $P_Y = 1$, the numeraire. The optimal rules for public assistance are as follows. First, have the rich buy food at the competitive price P_F so that $\text{MRS}^r_{Yr,\ \text{Fr}} = P_F$. Second, subsidize the food purchases of the poor person with a per-unit subsidy so that he buys food at the discounted price $P_F - s$, and allow him to consume as much food as he wants at the subsidized price. The poor person consumes Y and F such that his $\text{MRS}^p_{Yp,\ Fp} = P_F - \text{s}$. With $s = \sum_r \text{MRS}^r_{Yr,\ Fp}$, the Pigovian subsidy, the consumption of Y and F by rich and poor satisfies the pareto-optimal condition, Eqn (10.7). Finally, tax the rich in a lump-sum manner to pay for the food subsidies. Any pattern of lump-sum taxes on the rich maintains the pareto-optimal condition.

The tax-transfer policy of a Pigovian subsidy paid for with lump-sum taxes avoids the free-rider problem by forcing all the rich to participate in the program. Also, since this tax-subsidy policy moves society to the utility–possibilities frontier from somewhere under the frontier, there must exist a pattern of lump-sum taxes on the rich such that every rich person is better off with the policy. The increased utility to them of the poor person receiving more food exceeds the decreased utility from the taxes with the appropriate lump-sum taxes. The rich should not object to being forced to participate in an everyone-gains policy.

The model is easily extended to large numbers of poor and even different classes of the poor, say the near-poor (np), the poor (p), and the desperately poor (dp). As one possibility, assume that the utility of the rich is

$$U^r\left(Y_r, F_r, F_{np}, F_p, F_{dp}\right)$$

and add the assumption that every poor person within any one class is viewed identically by each rich person. The pareto-optimal policy now calls for three different Pigovian per-unit subsidies, with the subsidies presumably from the near-poor to the desperately poor.

What Motivates Charity: Should Aid Be In Kind or Cash?

The model we have been using calls for in-kind food subsidies to the poor because it is the consumption of food by the poor that concerns the altruistic rich. The Food Stamp program could be justified by this kind of model. Other in-kind public assistance programs such as Housing Assistance and Medicaid also suggest an underlying model of this form, with concerns about the housing and medical care of the poor added to the utility function of the rich. Over two-thirds of all public assistance in the United States is in kind, and the in-kind percentage has been steadily increasing over time and will continue to do so because of the rapid growth of Medicaid.

At the same time, however, approximately one-third of public assistance is in the form of cash, primarily monthly benefit checks, and cash was the principal means of supporting the poor when the federal government entered the public assistance arena during the Great Depression with the passage of the Social Security Act of 1935. The Act established three public assistance programs: Old Age Assistance, Aid to the Blind, and Aid to Dependent Children (later renamed Aid to Families with Dependent Children, AFDC). These programs gave monthly benefit checks to the poor who were also elderly, blind, or single parents (primarily widows in 1935). They also provided for payments to vendors of medical care to the recipients. Aid to the Disabled was added in 1951. In 1965, Medicaid consolidated all the medical vendor payments under the original programs and has since greatly expanded. In 1974, the federal government combined Old Age Assistance, Aid to the Blind, and Aid to the Disabled into one federal program, SSI. In 1996, the Congress replaced AFDC with TANF.

Explaining monthly benefit payments under the original public assistance programs, and now under SSI, TANF, and the EITC, requires a different model from the one described above. Cash assistance suggests a motivation in which the altruistic rich look at the poor and see that they are lacking all kinds of goods and services, not just food or housing or medical care. They decide that the poor need more income to reach even a minimally adequate standard of living, to be spent as the poor see fit. Returning to the simple two-person model, the utility function of the rich would include the entire utility function of the poor as one of its arguments:

$$U^r = U^r\left(Y_r, F_r, UP\left(Y_p, F_p\right)\right)$$

The utility of the rich person is greatest when the utility of the poor person is as high as possible for any given amount of aid, and the utility of the poor person is highest with a cash transfer in general, not an in-kind transfer.

Recipients' Preference for Cash

Figure 10.1 illustrates the advantage of a cash transfer from the vantage point of the poor recipient. It assumes that the market prices of F (Food) and the composite commodity Y are both equal to one. AB is the budget line of the poor person (P) given his own resources, without any transfer from the rich (R). P is initially in equilibrium at point E on AB and reaches the indifference curve I_0. A food subsidy rotates the budget line outward from point A to line AC. P achieves a new equilibrium at point M on budget line AC, reaching indifference curve I_1. At M, P spends GH of his own resources on F and receives a subsidy of HM. The percentage subsidy is (HM/GM) · 100.

Compare the in-kind food subsidy HM with a cash transfer equal to HM. The cash transfer causes the budget line AB to shift out parallel by an amount HM, to the new budget line JK. P achieves a new equilibrium at point N on budget line JK and reaches the indifference curve I_2.

In general, N contains more Y and less F than M and is on a higher indifference curve, as shown in the figure. The first point follows because the in-kind food subsidy generates a substitution effect in favor of purchasing F that is missing with the cash grant. They both have the same income effect, represented as the value of the transfer HM. The second point follows by a revealed preference argument. When P purchased the combination of Y and F at N with the cash subsidy, he could have purchased the combination at M; when he purchased the combination at M with the food subsidy, he could not have purchased the combination at N. Therefore, N is revealed preferred to M. Intuitively, P has to bias his purchases toward F to generate the transfer of HM under the in-kind food subsidy, whereas he receives HM under the cash transfer no matter what he buys. The subsidy acts as an additional constraint on P's options and lowers his utility relative to a cash transfer of equal value.

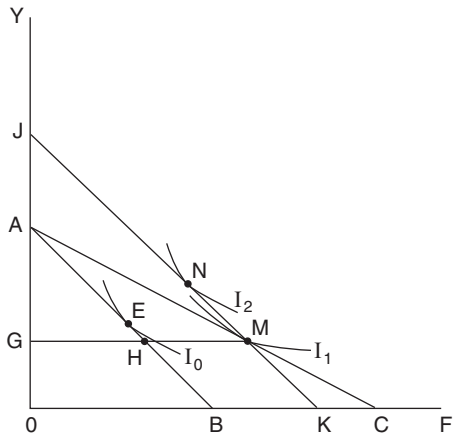

FIGURE 10.1

Another way to see that cash is preferred under the new model is to ask how the rich person responds to the poor person's purchases of Y and F. The relevant MRS from the point of view of the rich person is

$$\frac{\partial U^r}{\partial U^p}\frac{\partial U^p}{\partial Y_p}\Big/\frac{\partial U^r}{\partial U^p}\frac{\partial U^p}{\partial F_p} = \frac{\partial U^p}{\partial Y_p}\Big/\frac{\partial U^p}{\partial F_p} = \mathrm{MRS}^p_{Y_p,F_p} \quad (10.8)$$

the poor person's own MRS. Therefore, the utility of the rich is maximized if the poor person buys Y and F at the going market prices, implying that the transfer should be in cash.

The mixture of in-kind and cash public assistance in the United States gives mixed signals about how the nonpoor view the poor. The in-kind aid suggests that altruistic impulses are moderated by paternalism, that the nonpoor give in-kind aid of basic goods and services because they want accountability for their charity. They fear that the poor would tend to spend cash transfers irresponsibly, against the best interests of themselves and their families.[4] The cash transfers suggest a purer form of altruism, a willingness to extend the principle of consumer sovereignty to the poor. The nonpoor give cash because they believe that the poor, like themselves, are best able to judge their own self-interests and will spend any additional income they receive responsibly.[5] Do the nonpoor believe that the poor have fundamentally different preferences from them or that they simply have less income? The nonpoor in the United States have not given a clear answer to this question.

Limited Aid: Cash Equivalent In-Kind Transfers

The desire for accountability through in-kind transfers may be difficult to achieve if, as is often the case, the amount of aid per person or family is limited. The model above that justifies in-kind aid calls for unlimited subsidies of F: let the poor buy as much food as they want at the discounted price. Yet, governments almost always put limits on the amount of aid that can be received.

One reason for limiting aid is budgetary accountability. Refer again to Fig. 10.1. With the unlimited subsidy, legislators cannot know the amount of aid they will be giving

4. The professional literature has analyzed other nonaltruistic reasons for preferring in-kind aid that are based on imperfect information, such as the inability to monitor the behavior of the aid recipients. A world of imperfect information is inherently a second-best environment, so we will consider these other motives for in-kind aid in Chapter 19.

5. A decidedly less noble spin on the willingness to give cash has been suggested by Gordon Tullock. He argues that giving cash may be motivated out of fear, namely, that the poor will rise up against the nonpoor and try to seize their property. The nonpoor respond by trying to buy off the poor with aid, and the most effective way to do this is to maximize the satisfaction of the poor per dollar of aid. That is, give them cash. See Tullock (1983), Chapter 1.

until the poor make their spending decisions. In terms of the diagram, the amount of aid HM is unknown until the poor select point *M* on the subsidized budget line AC. Legislators do not like that kind of uncertainty so they place a limit on the amount of aid to have in advance a better sense what their commitment will be.

A second reason for limits is to avoid the possibility of resales. Under an unlimited-subsidy Food Stamp program, for example, the poor could buy the stamps at a given discount and resell the coupons to anyone at a slightly higher price, but one that is still well below the market price. The demand for food stamps would be unlimited, which is a powerful incentive for imposing limits on the amount of stamps any one person can receive. Housing assistance and Medicaid are less prone to resales than food stamps, but the desire for budgetary accountability still applies and leads to limits on these very expensive items.

The problem with placing a limit on in-kind aid is that it can make the in-kind aid equivalent to a cash transfer and undermine the nonpoor's desire for accountability. To see this refer to Fig. 10.2. The figure reproduces the same initial conditions as in Fig. 10.1. The budget line without aid is AB, and the poor person is initially in equilibrium at point E. The government offers a food subsidy at the same rate as in Fig. 10.1 (HM/GM · 100), but this time with a limit on the total amount of aid equal to RT. Once the limit is reached, the with-aid budget line continues parallel to AB at a horizontal distance RT beyond AB. The with-aid budget line is ATW. The poor person reaches a new equilibrium at point O on ATW. Suppose, instead, the government offered a cash transfer in amount RT. This would shift the budget line to XW, and the poor person would again reach a new equilibrium at point O, the same point as with the limited in-kind aid.

Limited in-kind aid is always equivalent to a cash transfer as long as the recipient spends more on the aided item than the total amount spent when the subsidy reaches its maximum. This amount is LT in Fig. 10.2, less than the amount of F purchased at O. Alternatively, in-kind aid is equivalent to cash if *marginal* purchases of the aided item occur at the full market price. This applies to virtually all families who receive food stamps, which is why economists view the Food Stamp program as just another cash transfer to the poor.

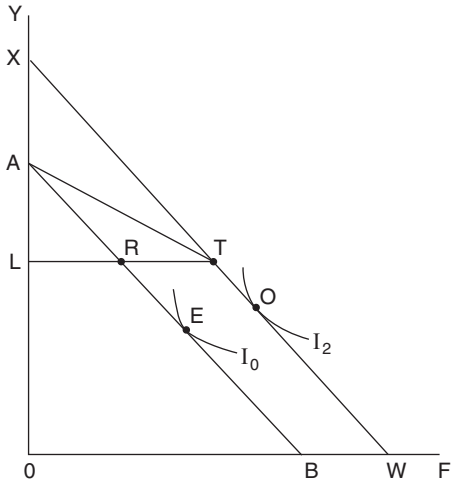

FIGURE 10.2

The intuition for cash equivalence is that the substitution effect under the subsidy program ends beyond point T, leaving only the same income effect as under a cash transfer. Therefore, recipients can undo the in-kind condition by reducing expenditures from their own incomes on the aided item until they reproduce what they would have done under an equal value cash grant. Put differently, limited in-kind aid differs from a cash transfer only if the recipient does not reach the limit—in terms of Fig. 10.2, if the recipient ends up somewhere on line segment AT under the in-kind program. Only then has the in-kind aid imposed some accountability on the poor by biasing their expenditures toward the aided item relative to a cash grant. The bias is due to the substitution effect, which does apply below the limit.

Are Pareto-Optimal Redistributions Enough?

Pareto-optimal redistributions cannot by themselves fully resolve society's quest for distributive justice, for end-results equity. They may be part of the recipe for the optimal distribution, but they cannot be the entire recipe. This is because a gain−gain redistribution motivated by altruism only serves to restrict the range of the first-best utility−possibilities frontier. The efficient pareto-optimal redistribution selects one point on the restricted frontier but, as with all efficiency rules, it cannot judge whether it has chosen the best point on the frontier. Selecting the first-best bliss point still requires a social welfare function.

To see this, refer to the utility−possibilities frontier in Fig. 10.3. Suppose the two people whose utilities are pictured in the figure are altruistic toward one another. Begin at point A, at which person 2 has everything. Because person 2 is altruistic, he is presumably willing to transfer some income to person 1. Therefore, both people gain from the transfer and the utility−possibilities frontier moves in a northeast direction from A. At some point though, say at point B, person 2 decides that person 1 has enough and is unwilling to transfer more income to her. Any further (forced) transfers are lose−gain propositions, and the utility frontier moves in the usual southeast direction from B. The same argument applies in reverse at point D, at which person 1 has everything. The utility−possibilities frontier moves in a northeast direction from D until some point C, when person 1's willingness to

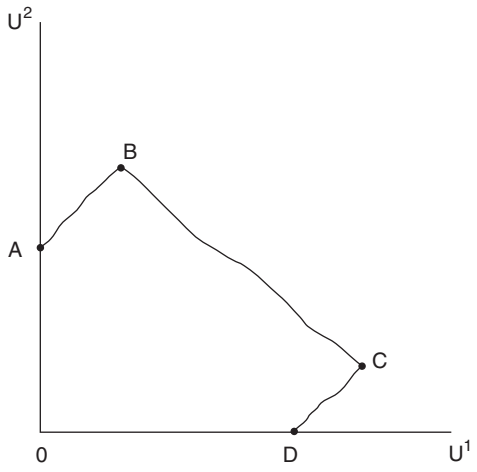

FIGURE 10.3

transfer to person 2 ends. Any further (forced) transfers from person 1 to 2 are lose−gain propositions, and the utility frontier moves in the usual northwest direction from C. Therefore, pareto-optimal redistribution restricts the utility−possibilities frontier to the line segment BC.[6]

If the economy begins to the left of point B or the right of point C, a pareto-optimal redistribution would bring the economy to B or C. The economy can achieve points between B and C on the frontier starting from other initial distributions, and a pareto-optimal redistribution may or may not be part of the complete set of pareto-optimal conditions on the interior segment.

Which is the best point, the bliss point, on the restricted frontier? Society cannot answer this question without recourse to the social welfare function. The pareto-optimal conditions are never sufficient by themselves to determine which of the efficient allocations is distributionally the best, even if the efficiency conditions themselves imply some redistribution motivated by altruism. It is always the interpersonal equity conditions from social welfare maximization that bring the economy to the bliss point.

Formally, the general equilibrium model described at the beginning of this section would have to be specified as a social welfare maximization in the usual manner to determine the first-best bliss point. The model would then describe two types of redistribution: one pareto-optimal redistribution, either cash or in kind depending on the nature of the altruism, and one set of lump-sum taxes and transfers of some good or factor to satisfy the interpersonal equity conditions. The two redistributions depend on one another and are determined simultaneously.

An argument can be made that one set of redistributions is enough (namely, the pareto-optimal redistribution), but the argument is not entirely convincing. It presumes, first of all, that society is willing to accept the initial distribution of resources whatever it may be. It also allows the nonpoor donors to have complete say over the amount of distribution that takes place; the poor are effectively disenfranchised in the quest for distributive justice. The poor in such a society may not fare very well if the initial distribution of resources is highly skewed and the nonpoor are not very charitable.

The social welfare function brings two distinct advantages to society's quest for end-results equity: it implicitly gives everyone a vote through the political process on the distribution question, and it adjusts for perceived inequities in the initial distribution of resources through the interpersonal equity conditions. In truth, the social welfare function is not so easy to discard from a normative theory of the public sector, however problematic it may be.[7]

ALTRUISM, FREE RIDING, AND CROWDING OUT OF PRIVATE CHARITY

The notion of pareto-optimal redistribution has practical as well as theoretical difficulties when private charity and public assistance exist side by side. Gain−gain redistributions motivated by altruism have two very strong properties. One is the powerful incentive for donors to free ride on the gifts of other donors. The other is that public assistance crowds out (reduces) private charity dollar for dollar under altruism. Neither property is even roughly consistent with the facts in the United States.

James Andreoni has developed a simple endowment model with altruism to illustrate the effects of these two properties for large economies. He begins with the case of only private charity and explores the propensity to free ride on the gifts of others (Andreoni, 1988).

Do People Free Ride?

Assume a nation of N people in which everyone has the same tastes, with utility defined over a composite commodity good y and the total amount of charitable giving, G:

$$U^i = U^i(y_i, G) i = 1, \ldots, N$$

$P_y = 1$, the numeraire, and a unit of G is \$1, a cash grant. Each person i has an endowment w_i.

6. This analysis appears in Boadway and Wildasin (1984).

7. Readers interested in models of altruism might consult Ley (1997). Ley cautions against the potential pitfalls of simple linear utility representations of altruism of the form $V_i = \beta_{ii}\, U_i + \sum_{j \neq i} \beta_{ij}\, U_j$, $\beta_{ij} \geq 0$. To give one example, he considers the case in which utility is a function of one private composite good and one public good and shows that all pareto-optimal allocations in the altruistic economy are pareto-optimal allocations in the egoistic economy in which utility is a function only of one's own consumption. The linear representation of altruism does not buy anything.

Define g_i as person i's own charitable contribution, and G_{-i} as the total charitable contributions of everyone except person i. Assume a Nash environment in which person i takes G_{-i} as given. Under the Nash assumption each person i solves the problem:

$$\max_{(y_i, g_i)} U^i(y_i, G) \quad \text{equivalently} \quad \max_{(y_i, G)} U^i(y_i, G)$$

$$s.t.\ y_i + g_i = w_i \qquad g_i \geq 0 \qquad s.t.\ y_i + G = w_i + G_{-i}, \quad G \geq G_{-i}$$

Using the equivalent formulation on the right, the demand for G can be written as

$$G = \max\{\gamma(w_i + G_{-i}), G_{-i}\} \quad i = 1, \ldots, N \tag{10.9}$$

where $\gamma(\ )$ is i's Engel curve for charitable giving, identical for all individuals. Assume that y and G are both normal goods, so that $0 < \gamma' = a < 1$. If person i is at an interior solution, then:

$$G = \gamma(w_i + G_{-1}) \tag{10.10}$$

Invert γ and then add g_i to both sides to obtain:

$$\gamma^{-1}(G) = w_i + G_{-1} \tag{10.11}$$

and

$$g_i = w_i + G - \gamma^{-1}(G) = w_i - \phi(G) \tag{10.12}$$

with

$$\phi(G) = \gamma^{-1}(G) - G \tag{10.13}$$

Note for future reference that $\phi' = 1/a - 1$, and $\phi^{-1'} = a/(1-a) < \infty$.

Let w^* = the amount of endowment at which the individual is just indifferent between giving and not giving. From Eqn (10.12)

$$g_i = 0 = w^* - \phi(G) \tag{10.14}$$

or

$$w^* = \phi(G) \tag{10.15}$$

Therefore, also from Eqn (10.12)

$$\begin{aligned} g_i &= w_i - w^*, \quad for \quad w_i \geq w^* \\ g_i &= 0, \quad for \quad w_i \leq w^* \end{aligned} \tag{10.16}$$

and

$$G = \sum_{w_i > w*} (w_i - w^*) \tag{10.17}$$

But, $G = \phi\ (w^*)$ from Eqn (10.15). Therefore,

$$\phi^{-1}(w^*) = \sum_{w_i > w^*} (w_i - w^*) \tag{10.18}$$

Next consider the average amount of charity per person, H_N, equal to

$$H_N = \phi^{-1}(w^*)/N = 1/N \sum_{w_i > w*} (w_i - w^*) \tag{10.19}$$

and ask what happens to the average as N becomes large.

Note, first, that the level of wealth at which an individual is just indifferent to giving varies with N. Thus, the general expression for the average amount of charity per person, H_N, is

$$H_N(s) = \phi^{-1}(s)/N = 1/N \sum_{w_i > s} (w_i - s) \tag{10.20}$$

As N becomes large without limit, total giving $G = \phi^{-1}$ (s) is bounded if wealth is bounded because $\phi^{-1'} = a/(1-a) < \infty$. Therefore: $H_N(s)$, the average gift per person, goes to 0.

To see what happens to the distribution of giving as N becomes large without limit, define an income distribution density function $f(w)$ over the continuum of individuals. The average gift per person is the expected value over the range of giving, or

$$\lim_{N \to \infty} H_N(s) = H(s) = \int_s^{\overline{w}} (w - s) f(w) dw \tag{10.21}$$

where $\overline{w}$ is maximum value of wealth in the economy.

But, the expected value is 0, so that the level of wealth, w^{**}, that divides those who give from those who do not give is the solution to the equation:

$$H(s) = \int_{w**}^{\overline{w}} (w - w^{**}) f(w) dw = 0 \tag{10.22}$$

The only solution to Eqn (10.22) is $w^{**} = \overline{w}$: only the wealthiest individuals give to private charity.

In conclusion, this simple model of altruistic behavior yields two very strong conclusions for large economies:

1. Although total giving, G, grows as the economy grows, the average gift per person goes to zero.
2. Only the wealthiest individuals give anything to private charity; the propensity to free ride is almost universal.

Neither of these conclusions is remotely close to the truth in the United States. Andreoni reports that about 85% of US households donate to private charities. The vast majority of people do not free ride on the gifts of others. Moreover, the average gift per household was $200 in 1971, with a range of $70 per person for those in the lowest fifth of the income distribution to $350 per person for those in the highest fifth of the income distribution. Pure altruism simply cannot explain the pattern of donations to private charity in the United States.

Does Public Assistance Crowd Out Private Giving?

To test the crowding out hypothesis, Andreoni posits a simple form of public assistance operating entirely through the tax system. Donors are subsidized at a rate s to give to charity, with the subsidies paid for by lump-sum taxes, τ, on each individual. Both the subsidy rate and the lump-sum tax can vary by individual. This form of assistance roughly imitates the subsidies to private donations under the federal personal income tax: taxpayers can deduct a portion of their private donations in computing their taxable income. The net contribution to public assistance by person i, a_i, is the difference between his lump-sum tax and the subsidy he receives on his private donations:

$$a_i = \tau_i - s_i g_i \tag{10.23}$$

The total amount of public assistance given to charity is

$$A = \sum_i (\tau_i - s_i g_i) \tag{10.24}$$

Each person's utility is now defined over the composite commodity y and the total amount of private plus public giving, $G + A$. $U^i = U^i(y_i, G + A)$.

Person i now solves the following problem:

$$\max_{(y_i, g_i)} U^i(y_i, G + A)$$
$$s.t.\ y_i + g_i + \tau_i - s_i g_i = w_i$$

Alternatively, define the total giving by person i, c_i, as the sum of her private and public giving:

$$c_i = g_i + a_i \tag{10.25}$$

Then the total giving for the entire economy is

$$C = \sum_i c_i$$

Define C_{-i} as the total giving by everyone except person i. Under the Nash assumption, an alternative formulation of the utility maximization problem is

$$\max_{(y_i, C)} U_i(y_i, C)$$
$$s.t.\ y_i + C = w_i + C_{-i}$$

Under the assumption that wealth after taxes, $w_i - \tau_i$, is always positive, this problem is identical in structure to the problem above with only private charity, with C replacing G. The economy achieves the same equilibrium and has the same strong free-riding properties.

Adding public assistance does not change the equilibrium because people can adjust their private giving to offset fully any changes in public assistance caused by changes in either taxes or the subsidy rate. Totally differentiate Eqn (10.23) and set $da_i = -dg_i(dc_i = 0)$ to determine how g_i adjusts to hold total net giving constant for changes in τ_i and s_i. In other words, increases in public assistance crowd out private giving dollar for dollar under pure altruism.

Once again the facts are quite different. Andreoni reported that econometric estimates of the degree of crowding out of private giving by public assistance in the United States ranged from \$0.05 and \$0.28 per dollar of public assistance. A more recent estimate by Donald Cox and George Jakubson is also within this range. They found the crowding out effect of public transfers on private transfers to be around \$0.12 on the dollar (Andreoni, 1995; Cox and Jakubson, 1995).

Andreoni speculates that other motives besides altruism drive donations to private charity, such as envy, sympathy, a sense of fairness, and a perceived duty to give. His preferred explanation for the large amount of private giving in the United States is what he calls a "warm glow" effect: people simply feel good about the act of giving to private charities, and the presence of public assistance cannot entirely undo this effect.

In conclusion, the public choice model of pareto-optimal redistribution motivated by altruism cannot be a complete model of the optimal distribution of income, either in theory or in practice. It does not remove the need for a social welfare function to answer the end-results equity question of distributive justice, and it cannot provide an explanation of the patterns of private or public charity in the United States. Nonetheless, the concept of a pareto-optimal redistribution is an important contribution to first-best distributional analysis. Charitable impulses that occur independently of any political process or social welfare function are an important phenomenon, and they do have the properties of a consumer externality.

OTHER MOTIVATIONS FOR REDISTRIBUTIVE TRANSFERS

We conclude the chapter with brief discussions of some other motivations for redistributive transfers that appear in the literature.

Public Insurance

Redistributive transfers motivated by a desire for income insurance are consistent with the public choice perspective. Buchanan argued in his Nobel address that the framers of a nation's constitution might permit redistributive public insurance programs such as social security pensions and unemployment insurance if they choose to view the future behind a veil of ignorance in which the future is truly uncertain (Buchanan, 1987). This vantage point raises the possibility that some of the framers or their descendants

may become impoverished, in which case allowing for redistributive public insurance can be viewed as self-serving.[8] The demand for public or social insurance is considered in Chapters 20 (medical care) and 21 (public pensions) of this text.

Public assistance can also have an insurance motive, at least partially. Thomas Husted attempted to distinguish between altruistic and insurance motives for public assistance in the United States on the basis of survey data collected as part of the 1982 American National Election Study (Husted, 1990). The participants were asked whether spending on food stamps and on AFDC was too much, about right, or too little. Husted hypothesized that the motives for food stamps were likely to be purely altruistic, with accountability. In contrast, the motives for AFDC were likely to be a mixture of altruism and insurance because the majority of the spells on AFDC are very short term, often only a month or two. Using econometric techniques suitable for survey responses, Husted obtained estimates that support his hypotheses. The demand for food stamps was uniformly upward sloping in income, whereas the demand for AFDC was U-shaped in income. An upward-sloping relationship between public assistance and income is consistent with an altruism motive. The inverse relationship between AFDC and income at the lower incomes is consistent with an insurance motive among the near-poor.

Husted's interpretation of the insurance motive is reasonable but open to question. One wonders how the poor and near-poor were able to muster support for insurance-based transfers as they tend to have little political influence. A possibility is that the rich also support public assistance at least partially for its insurance properties as Buchanan had suggested, but that the regression equation cannot separate insurance and altruistic motives among the rich.

Social Status

Amihai Glazer and Kai Konrad have raised the possibility that charity may be motivated by the donors' desire to achieve status among their peers rather than from any altruistic or warm glow feelings (Glazer and Konrad, 1996). Donors understand that income confers status and that a charitable gift acts as a signal of a person's income. The larger the gift, the larger the presumed income of the donor and the greater the status achieved.

Glazer and Konrad present a model in which individuals' utility is a function of their own consumption and their income as perceived by others, net of their charitable donation. Their perceived income is directly related to the size of their charitable donation. The model can explain a number of features of private charitable giving that pose difficulties for models based on altruism, in particular: why so many people give to charities, why the vast majority of gifts are not anonymous, and why, when charitable organizations report gifts in ranges such as $500–$999, the majority of gifts are bunched at or near the low end of the range.

Status seeking may well be an important motive for private charitable giving but it has difficulty explaining the tolerance for public transfers, which are necessarily anonymous. In any event, charitable gifts motivated by status seeking are obviously self-serving in the extreme.

Equal Access

Edgar Olsen and Diane Rogers have speculated that in-kind transfers may be motivated in part by the idea that people ought to have (approximately) equal access to certain social necessities such as medical care (Olsen and Rogers, 1991). The call for equal access falls more within the realm of process equity than end-results equity.

Complete equal access would require that each individual's purchase of the social good is subsidized such that everyone can afford the same maximum amount of the good. Think in terms of a two-good model: one the social good (necessity) and the other a composite commodity of all the other goods. Equal access for the social good implies that all budget lines are rotated by the subsidies such that they start at the same point on the social good axis.

Olsen and Rogers present a model of altruism in which each person's utility depends on his or her own consumption and a function defined over the maximum amount of the social good that each person can consume. The function is zero under equal access and causes reductions in utility that increase with the differences among individuals in their maximum possible consumption of the social good. The government policy is a combination of income subsidies and price subsidies for the social good. One of their central results is that all efficient allocations that are pareto superior to some initial allocation below the utility–possibilities frontier reduce the original inequality of access.

The Prospect of Upward Mobility Hypothesis

Having considered a number of possible motivations for redistribution, a puzzling question remains: Why is there

8. This motivation differs from the standard information problems of moral hazard and adverse selection that can undermine the formation of private insurance markets for some contingencies such as ill health and lead to a demand for public insurance. These information problems are second-best instances of market failure. Note that public insurance arising from poor information will also be redistributive, and in a particularly distressing fashion—from the well-behaved to the misbehaving. Simply bringing the insurance into the public sector does not eliminate the moral hazard incentives.

not more public redistribution in democratic societies? The poor and the middle class far outnumber the rich, and they have an obvious incentive to take from the rich for their own benefit. Why are they reluctant to vote to do so?

A common hypothesis for the United States, in line with the idea of the American dream of expanding opportunities, is that the nonrich expect their incomes, and those of their sons and daughters, to greatly improve in the future. This prospect of upward mobility (POUM) leads them to vote against redistributive public programs. If they vote for redistributive policies now, they assume those policies will remain in place in the future, and they would not want to be taxed in the future to redistribute to those now expected to be below them in the distribution.

Roland Benabou and Efe Ok have developed a model to formalize the POUM hypothesis. It is based on the notion that the expected evolution of income over time from one period to the next is concave, a common assumption in the social mobility literature. That is, current lower level incomes are expected to increase more rapidly over time than current higher level incomes. Concavity of the transition of income over time is consistent with diminishing returns to increasing skill levels, to offer one of a number of possible explanations.

Figure 10.4 illustrates. Current income Y is on the horizontal axis and the expected future income Y' is on the vertical axis. The function f is the transition function from current to future income. If is normalized in the figure such that the person with the mean level of income currently has the same income in the future. With this normalization, those with current incomes below the mean expect their (mean corrected) incomes to increase and those with incomes currently above the mean expect their (mean corrected) incomes to decrease. Since the lower incomes are rising more rapidly, the future mean level of income, $\mu_{Y'}$, falls. Income distributions are almost always positively skewed, with the mean income well above the median income. Given the concavity of the transition function, the expectation is that the distribution will become less positively skewed over time.

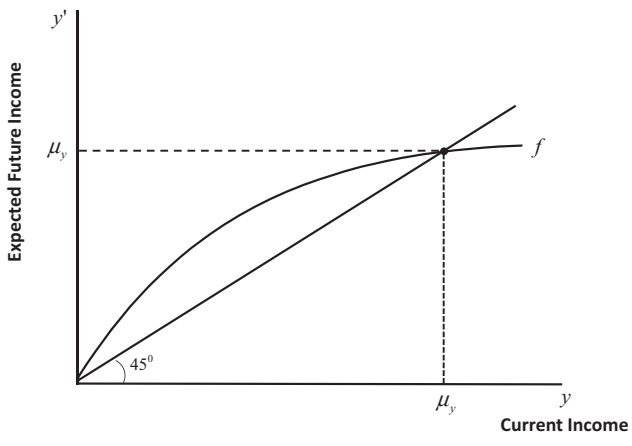

FIGURE 10.4

Given the expectations generated by the concave transition function, more people over time expect to be above the mean. Therefore, suppose for simplicity that people are asked to vote either for no redistribution (the status quo) or for the first-best policy of leveling all incomes to the mean. Faced with this choice, more people over time will vote for the status quo. If the transition function is highly concave, then it is possible that the expected future distribution would be negatively skewed, with the future expected mean income below the median income. In this case, a majority of the population votes in favor of the status quo. In fact, since low-income people tend to have the lowest voter participation rates, the future expected mean could still be above the median and a majority of voters would still support the status quo.[9]

The idea that the transition function is concave may well be unconvincing to people in the United States since so much of the growth in income since 2000 has gone to those in the upper tail of the distribution. As this is written, people in the United States talk about the death of the American dream, of POUM. But if this is now the future expectation, it may help to explain why the majority supported President Obama's call in 2013 for increasing personal income tax rates on the richest taxpayers to finance entitlement programs.

REFERENCES

Andreoni, J., February 1988. Privately provided public goods in a large economy: the limits of altruism. Journal of Public Economics 35 (1), 57–73.

Andreoni, J., February 1995. Warm glow vs. Cold prickly: the effects of positive and negative framing on cooperation in experiments. Quarterly Journal of Economics 110 (1), 1–22.

Benabou, R., Ok, E., May 2001. Social mobility and the demand for redistribution: the POUM hypothesis. The Quarterly Journal of Economics 116 (2), 447–487.

Boadway, R., Wildasin, D., 1984. Public Sector Economics, second ed. Little, Brown & Company, Boston, MA.

Buchanan, J., June 1987. The constitution of economic policy. American Economic Review 77 (3), 243–250.

Cox, D., Jakubson, G., May 1995. The connection between public transfers and private interfamily transfers. Journal of Public Economics 57 (1), 129–167.

9. Benabou and Ok (2001). Benabou and Ok offer a comprehensive formal analysis of the POUM hypothesis, adding uncertainty and income transitions over a number of periods.

Glazer, A., Konrad, K., September 1996. A signaling explanation for charity. American Economic Review 86 (4), 1019–1028.

Hochman, H., Rodgers, J., September 1969. Pareto-optimal redistributions. American Economic Review 59 (4:1), 542–557.

Husted, T., April 1990. Micro-based estimates of the demand for income-redistribution benefits. Public Finance Quarterly 18 (2), 157–181.

Ley, E., January 1997. Optimal Provision of public goods with altruistic individuals. Economic Letters 54 (1), 23–27.

Olsen, E., Rogers, D., June 1991. The welfare economics of equal access. Journal of Public Economics 45 (1), 91–105.

Tullock, G., 1983. The Economics of Income Redistribution. Kluwer-Nijhoff, Boston, MA.

Chapter 11

Applying First-Best Principles of Taxation—What to Tax and How

Chapter Outline

DESIGNING BROAD-BASED TAXES: THE ECONOMIC OBJECTIVES

Economists have proposed five economic objectives that governments should strive for in designing broad-based taxes:

1. Ease of administration and taxpayer compliance
2. Minimize deadweight loss
3. Promote long-run economic growth
4. Maintain flexibility
5. Honor society's norms of fairness or equity

The first objective takes precedence in the sense that if a tax does not meet both parts of this objective, it simply will not be used. Ease of administration refers to the ability of a department of revenue to collect the taxes due easily and economically, at a small fraction of the cost of the revenues raised. Ease of taxpayer compliance refers to the taxpayers'

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00011-X

ability to understand the tax code and pay the taxes owed with minimal effort, record keeping, and cost. The two are closely related, as taxpayers must be able and willing to pay their taxes for them to be collected easily. The need to satisfy the first objective explains why less-developed countries rely mostly on sales taxes, import duties, and other forms of business taxes rather than personal income and wealth taxes to raise revenue. Broad-based personal taxes such as income tax cannot be used if a large percentage of the population cannot read or write.

Objectives two and three refer to the efficiency properties of taxes: the second to static efficiency and the third to dynamic efficiency. Regarding static efficiency, we saw in Chapter 2 that buyers and sellers must face the same market prices to achieve the pareto-optimal conditions. Taxes distort markets by driving a wedge between the prices faced by buyers and sellers, thereby generating deadweight efficiency losses. The goal of tax design is to minimize the deadweight efficiency losses for any given amount of revenues collected. The dynamic efficiency problem is that taxes may also reduce incentives to save and invest, to the detriment of long-run economic growth. The goal is to maintain incentives for saving and investment to the fullest extent possible. A related problem is to ensure that tax policy keeps the economy as close as possible to the Golden Rule of Accumulation, the capital/labor ratio that maximizes consumption per person for any given rate of growth.

The flexibility objective is usually associated with the macroeconomic stabilization goal of smoothing the business cycle to keep the economy close to its production–possibilities frontier. Taxes are the main instrument of fiscal policy. As such, they must be flexible enough to be adjusted up or down as needed to smooth the business cycle.

The final objective calling for equity in taxation is a reminder that taxes must be consistent with society's norms in its quest for end results and process equity.

This chapter addresses only the final objective of achieving equity in taxation, for two reasons. One is that the pursuit of equity is a fundamental problem for a market economy that even a first-best perspective cannot assume away. The second is that the other tax design objectives are either less compelling or inapplicable in a first-best environment. The static and dynamic efficiency objectives, although very important to the design of broad-based taxes, are necessarily the second-best objectives and will be considered in Part III of the text. So, too, will the first objective. Ease of administration is generally not a serious issue for any of the broad-based taxes in the United States. All the major U.S. taxes are collected fairly easily and at very low cost. In contrast, taxpayer compliance is an important issue for some of the taxes, with the most serious problems resulting from private information. Taxpayers who are unwilling to pay their taxes may be able to hide information about themselves from the tax authorities. Private information is inherently a second-best issue, however. Finally, the macro flexibility issues are beyond the scope of this chapter.

We have seen in the previous chapters that first-best public sector theory does not provide much guidance to policymakers charged with designing broad-based taxes that the public will view as fair. The prescription for distributive equity in taxation (and transfer payments) is entirely contained within the interpersonal equity conditions for a social welfare maximum, yet these conditions beg the prior question of what the social welfare function should be. We also considered the benefits-received principle of taxation. It appears to have great appeal as a principle of tax equity in the United States, but its role in first-best theory is strictly as an efficiency principle. In any event, the benefits-received principle can only be narrowly applied to certain resource-using expenditures whose pattern of benefits is clearly defined. It cannot serve as the basis for designing broad-based taxes.

As it happens, tax practitioners have not been much bothered by the difficulties surrounding the social welfare function or the limitations of the benefits-received principle. Attempts to design fair broad-based taxes are almost always grounded in another principle of tax equity called the *ability-to-pay principle*, which dates from the beginnings of modern economics, having first been proposed by Adam Smith in the late 1700s and then further developed by John Stuart Mill in the early 1800s. The only established principle of tax equity before Smith and Mill was the benefits-received principle, which had originated in the fourteenth and fifteenth centuries under feudalism. The feudal lords would pay a tribute (tax) to the Crown in return for protection from foreign enemies. Smith and Mill recognized the need for another principle of tax equity for general taxes that were not so clearly tied to particular benefits received by the taxpayers.

The remainder of this chapter focuses on the ability-to-pay principle, indicating how to proceed from the principle to the design of broad-based taxes. The U.S. federal personal income tax will serve as the primary application throughout the chapter. Of all the broad-based taxes, it is this tax that is most closely grounded in the ability-to-pay principle.

The Smith–Mill ability-to-pay principle and the Bergson–Samuelson interpersonal equity conditions of first-best theory are also compared and contrasted. The older ability-to-pay principle would appear to bear a close kinship to the newer interpersonal equity conditions. The taxes and transfers implied by the interpersonal equity conditions surely depend on individuals' economic well-being, that is, on their ability to pay. Even so, the two principles are not as closely related as one might think.

They derive from fundamentally different views of taxation and, as such, they do not necessarily imply that the government should collect the same tax revenues from individuals or even use the same taxes.

ABILITY TO PAY: THEORETICAL CONSIDERATIONS

Smith and Mill recognized the limitations of the benefits-received principle as public expenditures became more varied and their benefits more diffused throughout the population. They reacted by introducing the concept of taxes as a necessary evil, a sacrifice that individuals have to make for the common good to support desired public expenditures. Given their perspective, they saw the fundamental question of tax equity as being one of how the government should ask people to sacrifice for the commonwealth, the common good. Their answer was that people should be asked to sacrifice in accordance with their ability to pay. In addition, the pattern of sacrifice should honor the two principles of *horizontal equity* and *vertical equity*. Horizontal equity says that equals should be treated equally. Two persons judged to have equal ability to pay should bear the same tax burden. Vertical equity allows for the unequal treatment of the unequals; that is, two persons with unequal abilities to pay can properly be asked to bear unequal tax burdens. This new Smith—Mill ability-to-pay principle was a sacrifice principle, pure and simple. Taxpayers should not expect a quid pro quo from general or broad-based taxes, in direct contrast to taxes paid according to benefits received.[1] Ability-to-pay principle was viewed as a default principle, to be used whenever the narrower benefits-received principle could not be applied.

The ability-to-pay principle quickly gained virtual unanimous acceptance as the appropriate equity norm for broad-based tax design. Its intellectual origins were familiar, dating from Aristotle and perhaps even further back, but a huge gap remained in applying the principles of horizontal and vertical equities to the actual design of a tax.

The requirements of horizontal and vertical equities beg two important and difficult questions. The first is the definition of equality: In what sense are two persons equal or unequal for the purposes of taxation? Both principles require an answer to this question. The second is the fundamental question in applying vertical equity: How unequally should unequals be treated under the tax laws? This is a part of the broader question of end-results equity, or distributive justice, related to the distribution of income.

The quest for horizontal equity in taxation has typically been associated with the goal of defining the *ideal tax base*. A person's tax liability is computed by multiplying the tax rate and the tax base. Therefore, two persons with the same value of the tax base necessarily pay the same tax and are treated equally in terms of taxation. The ideal tax base applies to vertical equity as well, since it defines the extent to which people are judged to be unequal for purposes of taxation.

Once the ideal tax base has been determined, the quest for vertical equity is then concerned with the design of the *tax structure*, which has two main components. First is the pattern of rates to be applied to different levels of the tax base. The second is the pattern of allowable exemptions, deductions, credits, and other adjustments to the tax base in computing the tax liability. These adjustments are justified in terms of promoting certain social goals that the government deems important. Two examples under the federal personal income tax are the personal exemptions that prevent the poor from having to pay taxes, which are permitted in the name of equity, and the deduction of interest payments on mortgages, which are permitted to encourage homeownership.

Two Preliminary Considerations

Two points should be noted before turning to the ideal tax base and tax structure. The first point is the fundamental difference in perspective between the Bergson—Samuelson interpersonal equity conditions of first-best theory and the Smith—Mill ability-to-pay principle. The taxes called for by the interpersonal equity conditions are inherently viewed as a good in and of themselves, since the interpersonal equity conditions are one of the two sets of first-order conditions necessary for maximizing social welfare. They promote social welfare by helping society reach the best distribution of income or utility on the utility—possibilities frontier. Taxes are not at all the necessary evil that Smith and Mill saw them to be. This sharp difference in perspective helps to explain why these two theoretical principles do not necessarily imply the same taxes, even if the taxes required by the interpersonal equity conditions are levied on the basis of ability to pay. We will return to this point after developing the implications of the ability-to-pay principle for the design of taxes.

The second point is that the ability-to-pay principle can properly be considered as a part of first-best theory. Ability-to-pay as a sacrifice principle relates specifically to the goal of distributive justice. Second-best tax theory is concerned, first and foremost, with the efficiency costs of distorting taxation. In a many-person second-best environment, efficiency considerations must be tempered by the equity implications of alternative distorting taxes, so that second-best theory has an interest in ability-to-pay principles. But the principles themselves have nothing whatsoever to do with the questions of efficiency. Hence, ability-to-pay principles

1. Smith (1904), Mill (1921). For an excellent history of the development of ability-to-pay principles, see Musgrave (1959), Chapter 5.

are analyzed most conveniently in a first-best environment, one in which efficiency and equity issues are separable. This is precisely what happened in the professional literature.

Careful distinctions between first-best and second-best analysis are a fairly recent phenomenon, but it is clear that early ability-to-pay theorists were implicitly assuming a first-best environment. We have two clues on this. The first is that the ability-to-pay literature generally ignores efficiency considerations altogether. This would be impossible in a second-best framework. The second is that ability-to-pay theory has traditionally equated tax payments and tax burdens. This, too, implies a first-best environment, for reasons that can only be sketched at this point in the text.

Tax incidence theory, the subject matter of Chapter 16, distinguishes between the burden of a tax (who sacrifices as a result of the tax) and the impact of a tax (who physically pays the tax—writes the check—to the government). We were careful earlier when defining horizontal and vertical equities to refer to "tax burdens." This is not always done. The two principles are often defined in terms of "tax payments" as follows:

- *Horizontal equity*: Equals should pay equal taxes.
- *Vertical equity*: Unequals should pay unequal taxes.

The difference is significant. Tax incidence theory shows that under certain conditions in a first-best policy environment, lump-sum tax payments are an appropriate measure of individual welfare losses, or burdens, using standard willingness-to-pay criteria such as Hicks' Compensating or Equivalent Variations. With distorting taxes, however, the tax payments are never entirely accurate measures of welfare loss. These points are fairly subtle and will be discussed in detail in Chapters 13 and 16. What matters here in terms of the ability-to-pay principles is that equal tax payments may yield unequal burdens with distorting taxes simply because of the distortions. Alternatively, unequal tax payments may entail equal burdens. Hence, once the possibility of distorting taxation is recognized, horizontal and vertical equities must be more broadly defined in terms of tax burdens, as we have done. Conversely, equating tax payments and tax burdens must imply both a first-best policy environment *and* lump-sum taxation.

We will adopt a first-best framework and equate tax payments and tax burdens to focus strictly on the equity issues involved with the ability-to-pay principles. This is at best an uneasy convenience, however. The problem is that the ability-to-pay principles lead to choices of broad-based taxes that are almost certainly not lump sum, so that it is impossible to ignore distortions entirely. In particular, the federal personal income tax contains a number of second-best distortions whose equity implications can only be understood in terms of the broader tax-burden interpretation of horizontal and vertical equities. Thus, we will occasionally stray from the first-best assumptions.

HORIZONTAL EQUITY

From Horizontal Equity to the Ideal Tax Base

Mainstream public sector economists do not agree on which tax base best satisfies the principle of horizontal equity. They do agree, however, on the proper way to think about what the ideal tax base should be. The line of reasoning from horizontal equity to the ideal tax base always relies on the same three principles of tax design. The disagreement occurs in applying the third principle, which describes the final step to the tax base.

The Three Principles of Tax Design

People Bear the Tax Burden

The first principle of tax design is that people ultimately bear the burden of any tax no matter what is actually taxed. For example, corporate income taxes and sales taxes are levied on business firms in the United States, but the fact that a business firm pays $X million in taxes is of little consequence. The interesting questions in terms of tax equity are which people finally bear the burden of these taxes. Is some or the entire burden "passed forward" to the consumers of the final product through higher prices, "passed back" to the labors employed by the firm through lower wages, borne by the stockholders of the firm, or borne by third parties not directly associated with the firm? Social well-being is directly related to individuals' utility functions, not to production relationships, and any tax eventually burdens people in their roles as consumers or as suppliers of factors, or both.

Individuals Sacrifice Utility

The second principle of tax design is that individuals ultimately sacrifice utility when they pay general taxes, so that the ideal tax base would be individual utility levels. In 1976, Martin Feldstein clarified what horizontal equity must mean to mainstream, neoclassical economists.

Feldstein's Horizontal Equity Principle: *Two people with the same utility before tax must have the same utility after tax.*

This is the only sensible economic interpretation of equal treatment of equals under a sacrifice principle of taxation.

Feldstein also proposed a minimum condition for the unequal treatment of unequals—no reversals—that has also gained universal acceptance among neoclassical economists.

Feldstein's Vertical Equity Principle (No Reversals): *If person i has greater utility than another person j before tax, then person i must have greater utility than person j after tax*[2]

Feldstein's two principles can only be guaranteed if utility is the tax base.

The Ideal Tax Base as the Best Surrogate Measure of Utility

Taxing utility is impossible, of course, but it still serves as a goal to strive for. Therefore, in lieu of taxing utility, the third principle of tax design is that the tax base should be the best practical surrogate measure of utility. Under this "ideal" tax base, the best surrogate for utility, two persons with equal values of the tax base are equals and should pay the same tax. This is as close as the tax practitioner can come to Feldstein's principle of equal utility before tax: equal utility after tax in the quest for horizontal equity.

Mainstream economists agree on the three principles, but they have not reached a consensus on what constitutes the best surrogate measure of utility. The two main contenders are income and consumption.

Haig–Simons Income

Neither Smith nor Mill was able to produce a convincing argument for an ideal tax base from their ability-to-pay principles. The first-proposed tax base that caught on appeared over 100 years later, in the 1920s and the 1930s. Robert Haig of Columbia and Herbert Simons of Chicago, following the line of reasoning above, independently concluded that a certain broad-based measure of income was the ideal tax base (Simons, 1938; Haig, 1921). Their proposal was almost universally adopted, and "Haig-Simons income" remained essentially unchallenged among mainstream economists as the best surrogate measure of utility until the 1960s, when consumption began to gain favor as a better surrogate measure. The majority of mainstream economists today may view consumption as the better choice.

Haig and Simons argued that purchasing power is the best surrogate measure of utility. This led them to propose income defined as the *increase in purchasing power during the year* as the ideal tax base for a tax levied annually. Using standard national income accounting terminology, Haig–Simons income can be defined as:

Haig–Simons income = consumption + the increase in net worth.

Consumption is the additional purchasing power actually taken, and the increase in net worth is additional potential purchasing power that has been deferred for future consumption. Net worth can be increased either by new saving or by an increase in the value of the individual's assets existing at the beginning of the year, the individual's *capital gains.* Therefore,

Haig–Simons income = consumption + saving + capital gains.

or

Haig–Simons income = personal income + capital gains[3]

Haig–Simons income is also called the *accretion standard* or, more commonly, the *comprehensive tax base*, a label so widely used now that it is often just referred to by the initials CTB.

Having determined that Haig–Simons income is the best surrogate measure of utility, horizontal equity is then defined as follows:

Horizontal equity: Two persons with identical amounts of Haig–Simons income are equals and should pay the same tax.

Similarly, two persons with different amounts of Haig–Simons income are unequals and should pay different taxes by the principle of vertical equity. The difference in their taxes depends on the tax structure applied to Haig–Simons income.

The Sources and Uses of Income

All components of Haig–Simons income are equivalent in terms of increasing purchasing power, so that the sources of income should not affect the amount of tax paid. The uses of the income are also irrelevant to the tax payment. Therefore, distinctions of the following kind should *not* matter in computing a person's tax liability, although they happen to matter under the U.S. federal personal income tax (violations of the Haig-Simons standard under the federal personal income tax are noted in brackets).[4]

Sources of Income

1. Whether income is derived from personal income or capital gains. (Capital gains are taxed at a substantially lower rate.)
2. Whether personal income is earned (wages, rents, etc.) or unearned (transfer payments). (Many transfer payments are untaxed, such as public assistance.)

2. Feldstein (1976). Feldstein's no-reversals principle is more than an equity principle. It also has important efficiency implications in a second-best world of imperfect information in which the government might not know how well off certain people are. Some people would have a powerful incentive to hide private information about themselves, if the tax laws permitted reversals of utility. We will return to this point in Chapter 15 when analyzing optimal second-best taxes.

3. Notice that the Haig–Simons definition uses personal income rather than disposable income because the former includes personal income taxes, which are originally part of the tax base.

4. The U.S. Internal Revenue Service refers to the tax as the individual income tax. Economists, however, typically refer to taxes levied on individuals as personal taxes and we adopt the economists' convention in this text.

3. Whether income is received in cash or in kind. (Many fringe benefits received by employees are untaxed, such as employer contributions to pensions and insurance.)
4. Whether earned income derives from labor, capital, or land. (Interest income on many forms of saving for retirement is exempt from income tax, such as the interest on Individual Retirement Accounts (IRAs).)

Uses of Income

1. Whether income is consumed or saved. Both consumption and saving increase utility. In terms of tax policy, the only relevant consideration is the increase in purchasing power, whether realized currently as consumption or postponed through saving. (Income used to purchase IRAs and some other retirement accounts is deductible from income in computing taxable income.)
2. Within capital gains, whether a gain is realized by selling an asset or simply accrues in value without a sale. Allowing gains to accrue is merely one particular form of saving. Also, capital losses should be fully offset against other income. (Capital gains are taxed only when realized, and there is only partial offset of capital losses.)
3. Consumption choices are also irrelevant, since all consumption decisions are viewed as voluntary and thereby utility increasing. These include contributions to private charities and tax payments to other governments to pay for the services they offer. (The following expenditures are deductible from income in computing taxable income (sometimes above some minimum level): medical expenses, contributions to charities and other nonprofit institutions such as colleges, state and local income and property taxes, and interest on a first mortgage.)

The only legitimate deduction from Haig–Simons income is the expenditures necessary for earning income in the first place, so-called *business expenses.* Presumably income used in this manner does not represent an increase in utility-enhancing purchasing power.[5]

Real versus Nominal Income

Haig–Simons income should be indexed for inflation so that inflation alone does not affect a taxpayer's real tax liability. Real income, not nominal income, is the better surrogate measure of the increase in purchasing power during the year. This point is important for an income tax since it taxes income from capital, which can differ greatly in real and nominal terms. Indexing for inflation matters for all sources of income when a tax uses a set of graduated rates that increase with income, as the federal personal income tax does. (The tax rates varied in seven steps from 10% to 39.6% in 2013). Inflation itself can move a taxpayer into a higher tax bracket and increase the real tax liability. (Only some components of the personal income tax, such as the personal exemptions and the income defining the tax rate brackets, are indexed for inflation.)

Other Tax Bases

A final point is that *all* tax bases other than Haig–Simons income are inappropriate because they are not the best surrogate measures of utility. These include: all broad-based taxes such as sales taxes, gift and estate (inheritance) taxes, and value-added taxes; selective excise taxes (except when required by the benefits-received principle); and taxes on specific sources of income, such as the payroll (Social Security) tax and the corporation income tax. Also inappropriate is taxing wealth in any form, such as local property taxes. The increase in purchasing power during the year, not accumulated purchasing power, is the appropriate annual tax base. The flaw with all these other taxes is that they cannot guarantee that two persons with the same Haig–Simons income before tax bear the same tax burden as required for horizontal equity. In fact, equals in terms of Haig–Simons income are very likely to be treated unequally under these other taxes.

Criticisms of Haig–Simons Income

Although Haig–Simons income is a reasonable choice for a tax base under the ability-to-pay principle, it could not be expected to gain unanimous acceptance among economists and policymakers, and it has not. Haig–Simons income is vulnerable to both negative and positive attacks. The negative attack is that Haig–Simons income may be a terrible surrogate measure of utility, in which case it loses its appeal as the ideal tax base. The positive attack is simply the belief that there is a better alternative to Haig–Simons income as the ideal tax base. The increasing support among neoclassical economists for consumption or expenditures as the ideal tax base is an argument of this kind. Finally, economists who do not accept the neoclassical perspective are likely to believe that some tax base other than Haig–Simons income is the better alternative.

A Flawed Surrogate Measure of Utility?

The negative view that Haig–Simons income may be a poor surrogate measure of utility is worth some discussion because the same argument can be applied to all proposed

5. The only issue is what constitutes a legitimate business expense, and this is often fought out in the courts. Purchase of a uniform required for work is deemed a legitimate business expense. Commuting expenses typically are not. They are considered part of the overall consumption package when people choose to live in a particular community.

tax bases under the ability-to-pay principle. Haig−Simons income does not necessarily suffer relative to other tax bases on these matters. We will simply use it to illustrate the nature of these attacks.

Haig−Simons income is a perfect surrogate measure of utility if people have the same tastes, abilities, and opportunities; otherwise, it may be a very poor surrogate. This is easily seen by means of the simple labor-leisure model in which people exchange hours of leisure for income at a constant hourly wage, w. The budget constraint is

$$Y = w(24 - \text{leisure})$$

where Y is income, w is the hourly wage, and there are 24 h in the day. Labor is the only source of income.

The two panels in Fig. 11.1 illustrate the difficulties with Haig−Simons income (wage income here) when tastes and opportunities differ. Tastes differ in the left-hand panel. One person is a leisure lover with indifference curves given by I_{LL}. The other person is a work lover (relatively speaking) with indifference curves given by I_{WL}. They face the same wage rate, w, the slope of the budget line. The diagram is meant to indicate that they have the same utility before tax because they reach the same numbered indifference curve, I^2. Therefore, they should have the same utility after tax by the principle of horizontal equity. But they have different incomes, Y_{LL} and Y_{WL}, so that they would pay different taxes with Haig−Simons income as the tax base. Consequently, their after-tax utilities may well differ, in violation of horizontal equity.

Opportunities differ in the right-hand panel. The two persons, 1 and 2, have the same tastes but face different wages, w_1 (the steeper slope) and w_2. The person facing the higher wage, w_1, is assumed to take all the additional purchasing power as increased leisure, to sharpen the point about income as a surrogate measure of utility. Person 1 is clearly better off, but they both earn the same income and therefore pay the same tax. Unequals are treated equally, in possible violation of both horizontal and vertical equities.

The failure of Haig−Simons income as a surrogate measure of utility in these examples is that it captures only one of the two variables that confer utility. The narrowness of income would not matter if the two persons were identical in every respect. It would then be a perfect surrogate for utility. These points are not peculiar to (Haig−Simons) income; they apply as well to anything chosen as the tax base. Income, consumption, or any component of income or consumption serves as a perfect surrogate for utility when people are identical in every respect, provided it is something purchased or earned by everyone (as opposed to an either-or choice of, say, a house or an apartment, which otherwise identical people may choose with indifference). Conversely, any one item that generates utility can be wide of the mark as a utility surrogate when tastes, abilities, and/or opportunities differ, because then all utility-generating items may matter in comparing utility.

A Better Alternative to Haig−Simons Income?

Is there a better alternative to Haig−Simons income as the ideal tax base for broad-based taxes? Many economists would say that there is.

To begin with, nonmainstream economists would not necessarily accept the three principles of tax design above as the path to the ideal tax base. Marxist economists, for example, would surely opt for differential treatment of wage and profit income for reasons that have nothing to do with surrogate measures of utility. As another example, Nicholas Kaldor is credited with the first serious proposal for a consumption or expenditures tax. He favored consumption not because of its relation with individual utility but from a broader social perspective. Kaldor agreed that consumption and saving are both self-serving choices by individuals designed to increase their

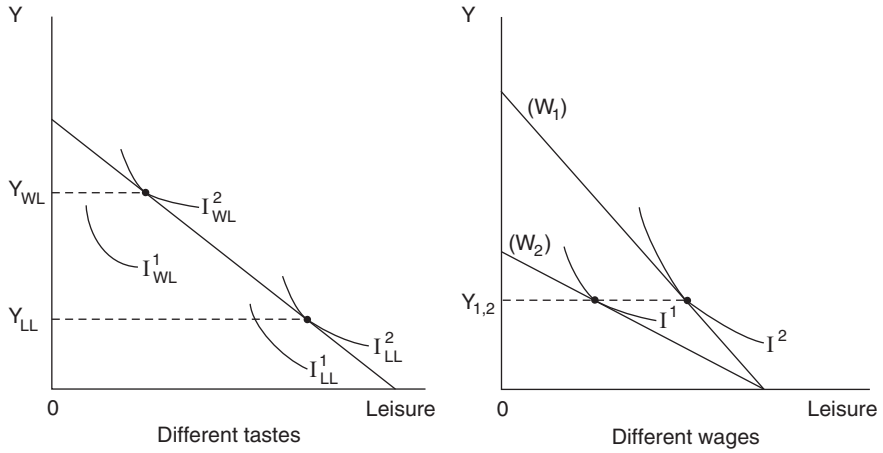

FIGURE 11.1

utility, either now or in the future, but he argued that society can meaningfully distinguish between the two, as follows. When individuals consume, they use up scarce resources for their own personal satisfaction, sacrificing others' well-being. In contrast, when individuals save, they provide funds for investment that leads to a more productive economy, to the potential future benefit of all citizens. Therefore, Kaldor (1955) argued that society can properly discriminate against consumption in taxation even if taxes are based on a sacrifice principle, providing sacrifice is viewed from a social rather than an individual perspective.

Consumption or Expenditures as the Preferred Alternative

The growing support among neoclassical economists for consumption or expenditures as the ideal tax base is in part based on Kaldor's argument. The only twist is that Kaldor's argument is seen today as a dynamic efficiency argument, not an equity argument. Simple, stylized, overlapping generations (OLG) models with perfect foresight that tracks the economy out for 100 periods and more find that replacing an income tax with a consumption tax leads to huge steady-state increases in output per person. Some models report increases on the order of 10–20%. The increased output results from the increase in saving, investment, and productivity under the consumption tax, exactly as Kaldor argued. This is seen as a powerful efficiency argument in favor of a consumption tax.

Many neoclassical economists add to this efficiency argument an equity argument that follows the standard three-step argument to an ideal tax base. They accept Feldstein's principle of horizontal equity—equal utility before tax, equal utility after tax—and the notion that the ideal tax base is the best surrogate measure of utility. But, they part company with the traditional Haig–Simons conclusion because they believe that the proponents of Haig–Simons income have the time frame wrong.

The break with the traditional view began in the 1960s following the development of Friedman's Permanent Income Hypothesis and Modigliani–Brumberg's life-cycle hypothesis (LCH), which themselves broke from the traditional Keynesian view of the consumption decision. The new theories viewed consumers as determining their consumption decisions over a longer period of time than a single year, indeed, over an entire lifetime in the case of the LCH.

The newer mainstream view of the ability-to-pay principle was that taxation should also be viewed in a lifetime context. Haig–Simons income is flawed as the ideal tax base because it relates only to a single year. Feldstein's equal utility before tax/equal utility after tax is the correct principle, but it should be applied to lifetime utility, appropriately discounted to present value: Two persons with equal present value of lifetime utility before tax should have equal present value of lifetime utility after tax. Therefore, the ideal tax base is the best surrogate measure of (discounted) lifetime utility.

The lifetime perspective argues for consumption, not income, as the ideal tax base by the following line of reasoning. The act of consumption is most closely related to the generation of utility. The Haig–Simons proponents have to think in terms of purchasing power because they adopt an annual perspective in which some purchasing power can be saved for future consumption. This is unnecessary in a lifetime perspective, however, because all income is eventually consumed (counting bequests to heirs as the final act of consumption). People receive income over their lifetimes in three forms: labor market earnings, inheritance, and other transfers from individuals and government.[6] They eventually consume all their income (again, counting the final bequest) such that the lifetime budget constraint holds: The present value of lifetime income equals the present value of lifetime consumption.[7]

From a lifetime perspective, therefore, the best surrogate for the present value of lifetime utility is the present value of lifetime consumption. Consequently, horizontal equity requires that two persons with identical present value of lifetime consumption before tax should have the same present value of lifetime consumption after tax. If taxes were levied on a lifetime basis, it would not matter whether consumption or income was the tax base, because the present value of lifetime consumption and income are equal. But taxes are levied on an annual basis, which matters. Only an *annual* consumption tax can guarantee that two persons with the same present value of lifetime consumption before tax have the same present value of lifetime consumption after tax.

An annual income tax breaks the equality between lifetime (discounted) consumption before and after tax because it effectively taxes saving twice. The income out of which the saving occurs is taxed, and any returns to the saving are also taxed. In other words, the pattern of consumption and saving matters in determining after-tax lifetime (discounted) consumption under an annual income tax, but not under an annual consumption tax. The following simple example illustrates this point.

6. Income from capital is not a source of lifetime income, at least not in an expected present value sense. Income from capital is expected to grow at the same rate of return, r, that is used as the discount rate to compute the present value of income. Therefore, any savings out of three sources of lifetime income cannot grow in expected present-value terms. Saving only changes the timing of consumption, not the overall present value of consumption, from a lifetime perspective.

7. In fact, most people lead virtually self-contained economic lives. The vast majority of people inherit very little wealth, which is the same as saying that most people bequeath very little wealth to their heirs.

Consumption versus Income Taxes: An Example

Suppose that two persons each live for two periods and earn a fixed amount of income Y in each period. Person 1 consumes the entire amount of income each period. Person 2 saves all of the first-period income and consumes everything in the second period. The savings earn a rate of interest, r, the same rate that they use to discount their second-period income and consumption to present value.

The top half of Table 11.1 gives the present value of lifetime consumption before tax, which is equal for both people. Under an annual consumption tax levied at rate t_c, the present value of lifetime taxes is the same for both of them: $\text{Taxes}_{PV} = t_c[Y + Y/(1 + r)]$. The only difference is that person 1 pays the tax in two installments and person 2 pays the tax all at once in the second period. Their present values of consumption are the same after tax, as required for horizontal equity.

The present value of taxes differs under an income tax at rate t_y, however, as illustrated by the bottom half of Table 11.1. Notice first, that the discount rate changes from $(1 + r)$ to $(1 + r(1 - t_y))$ under an income tax because interest income is taxed. The double taxation of saving occurs because the income of person 2 is taxed in the first period, so that only $Y(1 - t_y)$ is available as saving for second-period consumption, and then the interest on the saving is taxed again (assumed to be taxed in the second period here). The taxing of the interest income is what drives a wedge between the present value of taxes for the two persons. Horizontal equity is thus violated under an annual income tax. The two persons have equal present value of consumption before tax but unequal present value of consumption after tax.

The simple example also illustrates two ways to make an income tax equivalent to a consumption tax. One possibility is to allow taxpayers to deduct saving from income in computing their taxable income. This is an *expenditures tax*, which would be levied exactly as the personal income tax but with a deduction allowed for saving in computing taxable income. Since income is taxed only if consumed, an expenditures tax is the same as a consumption tax. In terms of the bottom half of Table 11.1, the deduction of saving removes the first-period tax from person 2 and also removes the tax on the interest income until it is consumed. With accumulating interest untaxed until consumed, the relevant discount factor reverts to $(1 + r)$, and the income tax with the savings deduction is fully equivalent to the consumption tax (assuming $t_y = t_c$).

The second possibility is to remove the double taxation of saving by allowing the taxpayer to deduct all interest income in computing taxable income (in general, any returns to saving/income from capital, whatever its form). This deduction also causes the discount rate to be $(1 + r)$ and removes the second tax term in second period for person 2. The income tax and consumption tax are once again equivalent.

Note, finally, that an expenditures tax is equivalent to a tax on wage income in this simple example because it is an income tax in which all income from capital is deductible. An expenditures tax and a wage tax are not equivalent in actual economies, however. The difference is that a wage tax is paid only during the working years, whereas an expenditures tax is paid in all years of life, including the retirement years. Neoclassical OLG models show that switching from a wage tax to an expenditures tax increases saving and investment because it hits the retired elderly particularly hard. They paid the wage tax while working and now they have to pay a tax on the consumption during retirement that they are financing from their accumulated savings while working. They also have the highest marginal propensity to consume of all the cohorts. The equivalent taxes in an OLG framework are an expenditures tax and a wage tax that includes a one-time capital levy on the retired elderly generation. The capital accumulated at the time of retirement equals the expected present value of

TABLE 11.1

	Period 1 Consumption	Period 2 Consumption	Present Value of Lifetime Consumption before Tax
Person 1	Y	Y	$Y + Y/(1 + r)$
Person 2	0	$Y(1 + r) + Y$	$[Y(1 + r) + Y]/(1 + r) = Y + Y/(1 + r)$
	Tax payments (income tax)		Present value of lifetime tax payments (income tax)
	Period 1		Period 2
Person 1	t_yY	t_yY	$t_y[Y + Y/(1 + r(1 - t_y))]$
Person 2	t_yY	$t_yY + t_yrY(1 - t_y)$	$t_y[Y + Y/(1 + r(1 - t_y)) + rY(1 - t_y)/(1 + r(1 - t_y))]$

consumption until death, counting any bequest as a final act of consumption.

The Tax Reform Act of 1986: Income Taxation versus Expenditures Taxation

The Tax Reform Act of 1986 (TRA86) was the largest single reform of the federal personal income tax ever undertaken, and the last reform of the tax base of any consequence. It made significant changes in the definition of taxable income and in the graduated rate structure applied to taxable income. The Reagan administration considered the possibility of replacing the income tax with an expenditures tax when preparing its initial proposal to Congress. The tax at that time was a mixture of the two kinds of taxes: essentially an income tax but with many features of an expenditures tax. The most important expenditures tax features were the treatment of various forms of pension savings such as IRAs and contributions to employer-sponsored pension plans. Contributions to these accounts and plans are deductible from income when made, and the accrued interest income until retirement is also excluded from taxable income. The pension incomes are taxed when received during retirement. This is exactly how savings of all kinds would be treated under an expenditures tax (provided that the pension income is consumed).

The administration decided to stay with the income tax, in large part because of the administrative headaches involved in switching from an income to an expenditures tax.[8] A particular sticking point was what to do about the elderly. They had already been double-taxed on their nonpension forms of saving. Under the income tax, they are not taxed again when they draw down their savings for consumption during retirement. If an expenditures tax were substituted for an income tax, the elderly would be taxed a third time as they consumed their savings. In truth, the large dynamic efficiency gains of switching from an income to an expenditures tax in an OLG framework come at an enormous cost to one group, the elderly, at the time of the switch. Burdening the elderly in this way was naturally considered grossly unfair, yet it was not clear how to protect the elderly (and near-elderly) during the changeover.[9]

8. The expenditures tax treatment of pension savings was retained, however, and still exists today.

9. The academic debate over income versus expenditures first heated up in the 1970s. See the articles by Richard Goode, David Bradford, and Michael Graetz in *What Should Be Taxed? Income or Expenditure*, (Pechman, 1980). Goode favors retaining the income tax, Bradford favors the expenditures tax, and Graetz offers an excellent discussion of the practical difficulties of changing from the income tax to an expenditures tax. See, also, Auerbach (2006).

Haig–Simons Income versus Expenditures: Musgrave's Perspective

Richard Musgrave believes that economists should call a halt to the income tax versus expenditures tax debate regarding horizontal equity. In his view, either Haig–Simons income or expenditures is an acceptable tax base. Neither one is a perfect surrogate measure of utility, but nothing else is either; and continuing to debate, which is the better utility surrogate, is pointless. Musgrave believes that vertical equity is far more important than horizontal equity in any event. Distributive justice is less affected by the choice of Haig–Simons income or expenditures as the tax base than by the tax structure applied to either.

According to Musgrave, the most useful way to interpret the call for horizontal equity in taxation is in a legalistic sense, the same way it is applied in other economic contexts. Equal treatment of equals should simply mean that the tax laws must never discriminate against people in inappropriate ways, such as on the basis of sex, race, or religion. Both Haig–Simons income and expenditures are admissible tax bases by this test. Therefore, Musgrave's position is that the federal government should simply choose one of them as the tax base and then worry about the appropriate tax structure (Musgrave, 1990).

Horizontal Equity and the Interpersonal Equity Conditions

Neoclassical economists would presumably want a tax designed in accordance with ability-to-pay principles to bear a fairly close relationship with the interpersonal equity conditions, since the interpersonal equity conditions are the ultimate guidelines for end-results equity in first-best public sector theory. Unfortunately, the ability-to-pay principle and the interpersonal equity conditions do not lead to the same pattern of taxation in general, even though the interpersonal equity conditions pay attention to peoples' economic circumstances, their ability to pay. The differences between them begin with the quest for horizontal equity.

Under the ability-to-pay principle, two persons are necessarily treated equally if they have the same tastes, abilities, and opportunities. Equal treatment under the interpersonal equity conditions also requires that people be identical over these three attributes but adds a fourth attribute as well: They must have the same marginal social welfare weights at equal levels of Haig–Simons income.[10] Two persons with equal utility before tax necessarily have

10. For the purposes of this discussion, assume that Haig–Simons income is chosen as the ideal tax base to satisfy horizontal equity and is the item redistributed lump sum to satisfy the interpersonal equity conditions.

equal utility after tax under the interpersonal equity conditions only if they are equal across all the four attributes. Furthermore, if two persons with equal utility before tax are not identical over the first three attributes, then they are not necessarily treated the same under the two principles even if they have equal marginal social welfare weights.

To illustrate, compare the interpersonal equity conditions and Feldstein's equal-utility-before-tax, equal-utility-after-tax criterion of horizontal equity within the context of a two-person, two-good exchange economy with fixed endowments of the two goods. Let X_{ij} = consumption of good j by person i, for $i, j = 1, 2$. The first-order conditions for a social welfare maximum in this economy are

$$\text{Pareto optimality:} \quad \frac{\frac{\partial U^1}{\partial X_{11}}}{\frac{\partial U^1}{\partial X_{12}}} = \frac{\frac{\partial U^2}{\partial X_{21}}}{\frac{\partial U^2}{\partial X_{22}}} \tag{11.1}$$

$$\text{Interpersonal equity:} \quad \frac{\partial W}{\partial U^1}\frac{\partial U^1}{\partial X_{11}} = \frac{\partial W}{\partial U^2}\frac{\partial U^2}{\partial X_{21}}$$

$$\frac{\partial W}{\partial U^1}\frac{\partial U^1}{\partial X_{12}} = \frac{\partial W}{\partial U^2}\frac{\partial U^2}{\partial X_{22}} \tag{11.2}$$

If the two otherwise-identical people have unequal social welfare weights, $\partial W/\partial U^1 \neq \partial W/\partial U^2$, evaluated at equal utility levels, the equal-utility-before-tax, equal-utility-after-tax criterion is inconsistent with Eqn (11.2), in general. This possibility can arise under an affirmative action policy that corrects for past injustices, as exists in the United States for women and minorities. The social welfare function can incorporate such a policy through the marginal social welfare weights, whereas the ability-to-pay principle cannot because it depends only on individuals' utilities.

Suppose the social welfare weights are equal so that Eqn (11.2) becomes

$$\frac{\partial U^1}{\partial X_{11}} = \frac{\partial U^2}{\partial X_{21}}$$

$$\frac{\partial U^1}{\partial X_{12}} = \frac{\partial U^2}{\partial X_{22}} \tag{11.3}$$

Even Eqn (11.3) differs from the horizontal equity criterion if people's tastes and/or initial endowments are unequal. If the two consumers happen to enjoy the same level of utility at an initial pareto optimum before the government redistributes one of the goods to satisfy the interpersonal equity conditions, there is no guarantee that they will enjoy equal utility levels after the socially optimum tax and transfer has been effected. The following simple model in which the two persons have different tastes provides a counterexample. Let

$$U^1 = X_{11}(3 + X_{12} + 27)$$

$$U^2 = \frac{1}{2}\cdot X_{21}\left(1 + 2X_{22}\right)$$

$$X_{11} + X_{12} = 10$$

$$X_{21} + X_{22} = 10$$

The reader can verify that an initial equal-utility pareto optimum occurs at

	$(X1, X2)$
Person 1	$(4, 2.4)$
Person 2	$(6, 7.6)$

The social welfare optimum, satisfying both Eqns (11.1) and (11.3), occurs at

	$(X1, X2)$
Person 1	$(5, 15/4)$
Person 2	$(5, 25/4)$

with unequal utilities. The difference occurs because the interpersonal equity requires equal after-tax *marginal* utilities, whereas the Feldstein criterion requires equal after-tax utility *levels*. Even after ignoring differences in social welfare weights, these two rules are consistent only if preferences and endowments are identical.

The unsettling conclusion is that the ability-to-pay principle of taxation is unlikely to be consistent with the interpersonal equity conditions of social welfare maximization. There are three differences between the two that cannot be fully reconciled.

The most important difference is that the interpersonal equity conditions add a new piece of information, the social welfare function, that is missing from the ability-to-pay principle. This alone is enough to generate inconsistencies between the two principles. The presence of the social welfare function also underscores their fundamentally different views of broad-based taxes: as promoters of social welfare on the one hand and as a necessary evil on the other hand.

A second difference is that horizontal equity under the ability-to-pay principle involves a before and after comparison of individuals' utility *levels:* equal utility before tax, equal utility after tax. The interpersonal equity conditions, in contrast, are concerned only with individuals' positions after tax (and transfer), and the comparison is in terms of margins, not levels: equal social *marginal* utilities. Moreover the equal-utility-level-before-tax, equal-utility-level-after-tax requirement is vulnerable as an equity principle because it is indifferent as to how two persons arrived at their equal utilities beforehand. One could be a respected

entrepreneur and the other a criminal, a distinction the social welfare function could take into account.

A final difference between them is that the quest for horizontal equity under the ability-to-pay principle is concerned with determining the ideal tax base, whereas the choice of the tax base is *irrelevant* under the interpersonal equity conditions. As we saw in Chapter 2, if pareto optimality holds and the interpersonal equity conditions are satisfied for any one good or factor, then the interpersonal equity conditions are automatically satisfied for all goods and factors, as required for a social welfare maximum. Any good or factor can be chosen for lump-sum redistribution; that is, *any* tax base will do. The only concern of the interpersonal equity conditions is vertical equity, the choice of the tax structure to be applied to whatever tax base is chosen. In summary, the ability-to-pay principles and the interpersonal equity conditions are quite different principles of taxation.

The question remains whether the ability-to-pay principle is a useful addition to neoclassical tax theory, given that the interpersonal equity conditions of social welfare maximization are *the* neoclassical statement of distributive equity. Might it not be better for policymakers to announce their preferred social welfare function, design a tax (and transfer) system that roughly corresponds to the requirements of the interpersonal equity conditions, and let citizens judge whether they are willing to accept the policymakers' social welfare function? What is gained by adding a completely different set of equity principles to the design of tax policy? These questions are in the spirit of Musgrave's suggestion to worry much more about the tax structure than the choice of an ideal tax base.

The practical answer appears to be that people are generally satisfied with the ability-to-pay principles. The Bergson–Samuelson social welfare function has had an enormous impact on the economic theory of the public sector but almost no impact at all on the design of broad-based taxes so far as equity itself is concerned. The only impact of social welfare analysis has been on the level of the tax rates, and then only when efficiency considerations are intermingled with equity considerations in a second-best environment. The interaction of efficiency and equity principles in taxation will be discussed in Chapters 14 and 15. The next step in this chapter is to consider the principle of vertical equity.

VERTICAL EQUITY

Once the ideal tax base has been determined, the quest for vertical equity centers on the design of the tax structure. Should the tax be levied at a single rate—a flat tax—or should the rates be graduated, rising with income? Should some minimum amount of income be exempt from taxation (assuming Haig–Simons income is the tax base)? Should taxpayers be allowed to deduct certain items of income or expenditure in computing their taxable income? The answers to these questions determine exactly how unequally unequals are treated under the tax laws, which is the central issue of vertical equity.

Progressive, Proportional, and Regressive Taxes

Actual policy discussions almost never get much further than the debate over whether taxes should be progressive, proportional, or regressive, three very broad indexes of vertical equity. Economists have devised various methods of defining these terms, but the most common definition is in terms of the average tax burden across individuals. Let

Y_i = value of the ideal tax base for individual i.

T_i = burden of the ideal tax on individual i.

The average tax burden on individual i is the ratio T_i/Y_i. Rank order individuals on the basis of Y_i and ask how the average tax burden varies as Y_i increases:

The tax *is progressive* if T_i/Y_i increases as Y_i increases.

The tax is *proportional* if T_i/Y_i remains constant as Y_i increases.

The tax is *regressive* if T_i/Y_i decreases as Y_i increases.

A number of points are worth stressing in applying this measure. The numerator should be the tax burden rather than the tax payment if the two differ, because the implicit standard is the relative loss in utility from the tax. By the same token, although the measure can be applied to any tax, the denominator should always be the ideal tax base for the purposes of assessing the vertical equity of the tax. The ideal tax base is the surrogate measure of an individual's utility and not anything else that might happen to be taxed. Additionally, the time frame should correspond to the time frame used to determine the ideal tax base. For example, proponents of Haig–Simons income as the ideal tax base should use it for the Y_i and the annual tax burden of a particular tax for the T_i. Proponents of consumption or expenditures should use the expected present value of lifetime consumption or income for the Y_i and the expected present value of the lifetime tax burden of a particular tax for the T_i.

A final point is that the three broad characterizations of vertical equity are not very limiting. Suppose, for example, that Haig–Simons income is chosen as the ideal tax base and society decides that it wants to collect more taxes from the rich than the poor under the ability-to-pay principle. A wide range of tax structures—progressive, regressive, or proportional—can satisfy the vertical equity criterion of unequal treatment of unequals and collect more taxes from higher-income individuals. For example, a tax structure that applies a 10% rate to an income of $50,000 and a 5% rate to an income of $200,000 is regressive. Yet, it collects more

tax from the richer individual, in broad concordance with the ability-to-pay principle.

About all one can say with confidence for the United States is that there appears to be an overwhelming consensus in favor of progressive or proportional taxes over regressive taxes. Studies of the overall U.S. tax system tend to show that the burden of all taxes is roughly proportional over all but the lowest income levels, within which they are slightly progressive. The U.S. tax system does not appear to redistribute much purchasing power in and of itself. We will return to this point in Chapter 17.

Vertical Equity and the Interpersonal Equity Conditions

In principle, the interpersonal equity conditions solve the problem of achieving vertical equity in tax design as part of determining the optimal distribution of income (assuming, again, that Haig−Simons income is the ideal tax base). Suppose that $Y^B = (Y_1^B, \ldots, Y_h^B, \ldots, Y_H^B)$ is the vector of Haig-Simons incomes across individuals before tax and transfer, and $Y^A = (Y_1^A, \ldots, Y_h^A, \ldots, Y_H^A)$ is the vector of Haig−Simons incomes across individuals after taxing and transferring to satisfy the interpersonal equity conditions. The difference between the corresponding elements in Y^B and Y^A defines the exact rate of tax (or transfer) to apply to each individual.

As usual, however, the interpersonal equity conditions are not very helpful to the tax practitioner. In addition to the uncertainties surrounding the social welfare function, the pattern of taxation may require that different tax rates be applied to people with the same Y_h^B if, say, the social welfare function incorporates a policy of affirmative action. Taxing different people differently on some basis other than their incomes may well be illegal in the United States. It also violates Musgrave's interpretation of horizontal equity as a proscription against taxation on the basis of inappropriate personal characteristics, a compelling proscription in matters of taxation.

Finally, we saw in Chapter 4 that attempts to apply the social welfare function typically assume: (1) equal marginal social welfare weights at equal incomes; (2) everyone has the same tastes; and (3) diminishing private marginal utility of income. The implication of the interpersonal equity conditions under these three assumptions is that everyone should have the mean level of income after tax and transfer. Hardly anyone accepts this view of vertical equity, perhaps because it is so difficult to ignore the efficiency implications of leveling everyone's income to the mean. And, indeed, the mainstream position is that the efficiency implications of any tax should be incorporated into a social-welfare-maximizing framework to determine the optimal structure of the tax.

Sacrifice Principles of Vertical Equity

Public sector economists had long worked on the problem of vertical equity from the sacrifice perspective of the ability-to-pay principle, but without much success until 1988. This line of research had pretty much died out by the 1980s. The main suggestions for vertical equity in the tax literature at that time dated from the late 1800s to the early 1900s. Then, in 1988, H. Peyton Young achieved a substantial breakthrough. Building on one of the earlier principles, Young used the methods of cooperative game theory to develop specific recommendations for the tax structure. Young's game-theoretic approach appears to be a promising avenue for future research.[11]

The two long-standing principles of vertical equity in taxation before Young wrote were minimum aggregate sacrifice and equal sacrifice.

Minimize Aggregate Sacrifice

The call to minimize the aggregate sacrifice from taxation came from the utilitarian school led by Jeremy Bentham, who believed that the economic goal of society should be to maximize aggregate happiness or utility. Their social welfare function was the straight sum of individual utilities. With broad-based taxes viewed as a necessary sacrifice for the common good, the corresponding utilitarian goal for tax policy was to minimize the aggregate sacrifice from collecting the taxes. Under the assumptions of identical tastes and diminishing marginal utility of income, aggregate sacrifice is minimized by levying taxes in a top−down, highly progressive manner until the required total tax revenue is collected.

To see this, suppose there are three groups of consumers having pretax incomes Y_1, Y_2, and Y_3, such that $Y_1 < Y_2 < Y_3$. Incomes are equal within each group. Assume further that their pretax marginal utilities of income are, respectively,

$$\frac{\partial U^1}{\partial Y^1} = 10 \quad \frac{\partial U^2}{\partial Y^2} = 9 \quad \frac{\partial U^3}{\partial Y^3} = 8$$

reflecting diminishing marginal utility.

If the government wants to collect a given amount of tax revenue, the minimum aggregate sacrifice principle requires that the government tax people in the third group until either their marginal utility rises to nine or the required tax revenue has been collected. If the former applies, then the government taxes both the second and third groups until either their marginal utility rises to 10 or the required tax revenue has been collected. If still more revenue needs to be collected, then the government taxes all three groups, maintaining equality on the margin, until the revenue

11. Young (1988). His companion in empirical exercise related to the U.S. personal income tax is Young (1990).

requirement has been met. This pattern of tax collections is highly progressive in terms of the tax burdens.

Equal Sacrifice

The other main suggestion called for equal sacrifice in terms of utility, the only debate being whether the government should require equal absolute sacrifice or equal proportional sacrifice. Letting Y_h be pretax income and T_h be the tax for person h, the two candidates are

Equal absolute sacrifice:

$$U(Y_h) - U(Y_h - T_h) = c \quad \text{all } h = 1, \ldots, H$$

Equal proportional sacrifice:

$$[U(Y_h) - U(Y_h - T_h)]/U(Y_h - T_h) = k$$
$$\text{all } h = 1, \ldots, H$$

The equal-proportional-sacrifice variation was a modern restatement of Aristotle's belief that proportional taxation was the just way to raise tax revenues.

Neither the utilitarian nor equal-sacrifice versions of vertical equity ever gained much standing among economists as a prescription for the design of a tax structure. One problem at the outset was the cardinality of the measures. The utilitarian prescription relies on diminishing marginal utility, which is neither a necessary nor sufficient condition for diminishing marginal rates of substitution, the condition for a well-behaved consumer indifference map.

Even if marginal utility is diminishing with respect to one utility index, there exists an admissible monotonic transformation of the utility function that leaves demands (and factor supplies) unchanged and implies either constant or increasing marginal utility. That is, given a utility index, $\phi(X)$ and its transformation, $F[\phi(X)]$, $F' > 0$,

$$\frac{\partial^2 F[\phi(X)]}{\partial X_i^2} = F' \frac{\partial^2 \phi}{\partial X_1^2} + \left(\frac{\partial \phi}{\partial X}\right)^2 F'' \geq 0 \qquad (11.4)$$

is consistent with $\partial^2\phi/\partial X_i^2 < 0$ for $F'' > 0$.

The same problem plagues the equal sacrifice principles. Equal absolute or proportional sacrifice with respect to $\phi(X)$ does not necessarily imply equal absolute or proportional sacrifice with respect to $F[\phi(X)]$. Needless to say, economists are skeptical of any economic principles based on cardinal utility measures.

Finally, suppose the government picked one cardinal representation of the utility index that satisfies diminishing marginal utility for each person in order to design a tax structure. Unfortunately, the pattern of taxes implied by any of the sacrifice principles could be just about anything. Even the utilitarian tax program need not be progressive. Using a simple general equilibrium model with one good and one factor, Efriam Sadka was able to show that lump-sum taxes consistent with the utilitarian social welfare function would not necessarily be progressive, where factor income is used as the basis of comparison. Whether the taxes are progressive or not turns on a number of parameters, including the elasticity of the consumers' indifference curves between the factor and the good, third derivatives of the utility function, and the like. Certainly no conclusions can be drawn a priori (Sadka, 1976).

Young's Prescription for Vertical Equity

H. Peyton Young revived the equal sacrifice ability-to-pay principle of vertical equity by introducing a new and thoroughly modern view of the problem of tax design. Young reasoned that if society views broad-based taxes as a necessary evil, a sacrifice made for the common good, then the levying of these taxes ought to be viewed as a cooperative game played by all members of the society. The design of the tax structure becomes the standard exercise in cooperative game theory of establishing a set of sharing rules for splitting up the profits or costs of the game. In this instance, the design problem is to posit a set of sacrifice principles that society could agree to in the levying of a broad-based tax and see what the principles imply for the tax structure. Arrow used the same cooperative game theory approach in proving his General Impossibility Theorem for social decisions in a democratic society.

Young posited six principles that he thought a democratic society could agree to in the levying of a broad-based tax. He then proved that they imply equal sacrifice in terms of one of two utility functions commonly used in the theory of risk taking. They also imply very simple tax systems.

We will assume that Haig–Simons income has been chosen as the tax base in demonstrating his result. Also everyone is assumed to have the same tastes; individuals vary only in the amount of income they have. We saw that the same-tastes assumption is necessary when selecting a tax base as a surrogate measure of utility. It is also necessary in order to say anything definite about vertical equity.

Young's Six Principles of Taxation

Young proposed the following six principles as the base for an equitable tax structure:

1. *The consistency principle*—If a method of taxation is considered to be fair for the entire group of taxpayers, then it must also be considered fair for any subgroup of the taxpayers. The force of this principle is to ensure that people cannot alter their tax liabilities simply by joining different subgroups. As such, it satisfies the requirement of coalition stability for solutions of cooperative games. The consistency principle is automatically satisfied if the tax is levied on individuals, since

different subgroupings or coalitions of taxpayers cannot possibly alter individual tax liabilities.[12]

2. *Monotonicity*—If the government is forced to increase total tax revenues, then everyone's tax liability must increase. This is the strong version of the principle. The weak version is that if total tax revenues increase, then no individual's tax liability can decrease. The monotonicity principle captures the spirit of ability to pay as a sacrifice principle, namely that the taxpayers are all in this game together. Notice that the strong version might not be satisfied by the utilitarian aggregate minimum sacrifice principle with its highly progressive, top−down tax collections.
3. *The composition principle*—The method used to raise a given amount of tax revenue must also be used to raise any increment in tax revenue. In other words, society should stick with the method that it believes is fair. This principle is satisfied by surtaxes, which raise additional revenue by requiring taxpayers to pay an additional percentage of their existing tax liability.

Feldstein's principles of horizontal and vertical equities constitute (4) and (5):

4. *Horizontal equity*—Two persons with equal utility before tax should have equal utility after tax.
5. *Vertical equity*−No utility reversals. For any two persons, the one with higher utility before tax must have higher utility after tax.

These two principles can also be stated in terms of Haig−Simons income since it is assumed to be an appropriate surrogate measure of utility.

6. *Scale invariance or the homogeneity principle*—Suppose everyone's incomes and the revenue requirement increase by a scalar 0. Then, everyone's tax liability must increase by θ. This principle is standard in income distribution theory, where it is applied to measures of income inequality. The idea is that an index of inequality should be invariant to scalar increases or decreases in everyone's income. It applies to relative tax burdens in this context.

The results of cooperative game theory rely on accepting the underlying principles, which may or may not be persuasive. If a democratic society were to accept Young's six principles of tax design, however, then the results are rather striking. Young proved that the first five principles hold if and only if the tax collections imply equal sacrifice with respect to some utility function, without specifying what that function should be. By adding the homogeneity principle, Young's six principles hold if and only if tax collections imply equal sacrifice with respect to one of the following two utility functions:

$$U^h = a\ln(Y_h) + b \quad \text{or}$$
$$U^h = aY^P + b \quad a, P < 0$$

These are the utility functions commonly used in the theory of risk taking because they exhibit constant relative risk aversion (CRRA), meaning that the elasticity of marginal utility with respect to income is constant (as the reader can easily verify). Equal sacrifice under these two utility functions in turn implies very simple tax functions: the first, a proportional tax and the second, a progressive tax.

An important point to note before demonstrating Young's results is that the distinction between equal absolute sacrifice and equal proportional sacrifice is irrelevant to modern economic theory. The reason is that equal absolute sacrifice with respect to some utility function, say U, is equivalent to equal proportional sacrifice with respect to the function e^U, which is a valid monotonic transformation of U and would have no effect on individual choice. To see this, assume equal absolute sacrifice exists with respect to U, such that $U(Y_h) - U(Y_h - T_h) = C$. Equal proportional sacrifice with respect to e^U is

$$\left[e^{U(Yh)} - e^{U(Yh-Th)}\right] \Big/ e^{U(Yh-Th)} = K \tag{11.5}$$

Simplifying Eqn (11.5) and rearranging terms, equal proportional sacrifice implies

$$e^{[U(Y_h)-(UY_h-T_h)]} = K = 1 = K' \tag{11.6}$$

which can only hold if $U(Y_h) - U(Y_h - T_h)$ is constant.

We will consider the sufficient conditions to see what Young's principles imply for the tax structure.[13] The first task is to show that each of Young's first five principles hold if the tax collections satisfy equal sacrifice with respect to some utility function. (Equal absolute sacrifice is easier to work with.) Therefore, suppose $U(Y_h) - U(Y_h - T_h) = C$, for $h = 1,\ldots, H$, and consider each of the first five principles.

1. *Consistency*—This holds by definition assuming that the tax base is each individual's Haig−Simons income.
2. *Monotonicity*—The strong version of monotonicity must hold under equal absolute sacrifice assuming positive marginal utility of income. Let total tax collections rise and assume that person i is taxed more. Then $U(Y_i) - U(Y_i - T_i) > C$. To maintain equal absolute sacrifice, everyone else must pay more tax to increase their

12. It is not satisfied by the U.S. federal personal income tax, however, because the IRS cannot decide if it wants to tax on an individual or a family basis. As a result, taxpayers within a family have the options of filing as individuals or pooling their incomes and filing jointly as members of a family. The individual and joint filing income cut-offs at which the different graduated rates apply differ, which means that taxpayers' liabilities can vary if they marry or divorce. Young's principle would permit only individual filing and thereby avoids the marry/divorce problem.

13. The necessary conditions are much more difficult to prove and will be left to a reading of Young's paper.

difference between $U(Y_h)$ and $U(Y_h - T_h)$ and restore equal sacrifice.

3. *Composition*—Assume that $U(Y_h) - U(Y_h - T_{1h}) = C$ for given total tax collections T_1. Suppose that tax collections rise to T_2 and equal absolute sacrifice is maintained for the increment of taxes between T_1 and T_2: $U(Y_h - T_{1h}) - U(Y_h - T_{1h} - T_{2h}) = C'$. Adding the two results: $U(Y_h) - U(Y_h - T_{1h} - T_{2h}) = C + C' = C''$. Equal absolute sacrifice is also maintained for the new higher tax collections T_2.
4. *Feldstein's horizontal equity principle*–Two people with equal utility before tax should have equal utility after tax.
5. *Feldstein's vertical equity principle*–No utility reversals. For any two people, the person with higher utility before tax must have higher utility after tax.

These two principles must hold under equal absolute sacrifice as long as the marginal utility of income is positive. Regarding horizontal equity, if $U(Y_i) = U(Y_j)$ and $U(Y_i) - U(Y_i - T_i) = C = U(Y_j) - U(Y_j) - (T_j)$, then $U(Y_i - T_i) = U(Y_j - T_j)$. Regarding the principle of no reversals, if $U(Y_i) > U(Y_j)$ and $U(Y_i) - U(Y_i - T_i) = C = U(Y_j) - U(Y_j - T_j)$, then $U(Y_i - T_i) > U(Y_j - T_j)$. Therefore, equal absolute sacrifice with respect to any valid utility function U satisfies each of Young's first five principles of taxation.

Now add the scale invariance or homogeneity principle, which generates Young's two proposed tax structures. The sufficient conditions on the tax structures involve two steps. First, determine the tax structure implied by equal absolute sacrifice with respect to the two CRRA utility functions noted above; second, show that the tax structures are scale invariant.

Proportional Taxation

Consider the utility function $U^h = a\ln\ Y_h + b$. Equal absolute sacrifice implies $a\ln\ Y_h - a\ln(Y_h - T_h) = C$, for $h = 1,\ldots, H$. The left-hand side is constant at any income if (and only if) $T_h = tY_h$, that is, under a flat-rate, proportional tax:

$$a[\ln Y_h - \ln(1-t)Y_h] = a\ln(1-t) = C \qquad (11.7)$$

A proportional tax clearly satisfies the homogeneity principle; the ratio $\theta T_h/\theta Y_h$ is independent of θ. Young's six principles of taxation have resurrected Aristotle's call for proportional taxation, assuming loglinear utility.

Progressive Taxation

Now consider the utility function $U^h = aY_h^p + b$. Equal absolute sacrifice implies

$$a\ Y_h^p - a(Y_h - T_h)^P = C \quad h = 1,\ldots,H \qquad (11.8)$$

Rearranging terms and solving for T_h yields

$$a\ Y_h^p - (Y_h - T_h)^P = C/a = -\lambda, \quad \text{with } a < 0 \qquad (11.9)$$

$$(Y_h - T_h)^P = (Y_h^P + \lambda) \qquad (11.10)$$

$$T_h = Y_h - (Y_h^p + \lambda)^{1/P} \qquad (11.11)$$

Under this tax, individual tax collections can be multiplied by a scalar as needed for total revenues. Therefore, the tax is a flat-rate tax applied to a tax base in which taxpayers exempt an amount $(Y_h^P + \lambda)^{1/P}$ from their Haig–Simons income (Y_h) in determining their taxable income. The tax has a number of interesting properties.

First, T_h/Y_h is independent of θ. This follows from dividing Eqn (11.11) by Y_h, and noting from Eqn (11.9) that scaling T_h and Y_h by θ scales λ by θ^P.

Second, the tax is progressive in terms of the standard average tax burden measure of progressivity. The average tax burden increases as Y_h increases (divide Eqn (11.11) by Y_h and recall that $P < 0$ and $\lambda > 0$).

Third, and most unusual, the exemption from the income in computing taxable income, $(Y_h^P + \lambda)^{1/P}$ *increases* as income increases. In all actual taxes with exemptions, the exemption either remains constant or decreases as income increases. Even so, the increasing exemption does not prevent the tax from being progressive.

Fourth, the homogeneity principle rules out graduated tax rates (although not progressive taxes).

In conclusion, Young has provided a rationale for either proportional or progressive broad-based taxes using the methods of cooperative game theory. In doing so, he has brought the old equal-sacrifice principle of taxation into the realm of modern economic theory. Whether he has done so successfully depends on a society's willingness to accept his six principles of fair taxation. Perhaps some other set of sacrifice principles would be viewed as more attractive and imply quite different tax structures.

Vertical Equity in the United States

The five major broad-based taxes in the United States give a mixed reading on how unequally the United States is willing to treat unequals. As we will see in Chapter 17, some of the taxes are progressive and others regressive. One can argue that the federal personal income tax gives the clearest signal of the U.S. view of vertical equity since it is designed on ability-to-pay principles. Unfortunately, it gives mixed signals as well.

The federal personal income tax appears to be fairly progressive on paper, with a graduated rate structure ranging from 10% to 39.6% and a large exemption of the first dollars of income to protect the poor from taxation. It

turns out to be much less progressive in practice, however, because capital gains and some other forms of income from capital receive highly favorable tax treatment, in some cases no tax at all. Capital income is highly concentrated among the richer taxpayers.

The recent history of the federal personal income tax has not clarified matters. TRA86 reduced the graduated rate schedule from 14 brackets ranging from 11% to 50% to five brackets ranging from 11% to 38.5% in 1987 and then to two brackets in 1988, 15% and 28%. The reduction of the top rate to 28% was done in large part to improve the dynamic efficiency of the tax. At the same time, TRA86 sharply increased the personal exemption to protect the poor from taxation. The history of the tax rates since 1988 has been marked by frequent changes: 1991, three brackets ranging from 15% to 31%; 1993, five brackets ranging from 15% to 39.6%; 2002, six brackets ranging from 10% to 38.5%; 2003, six brackets ranging from 10% to 35%; and 2013, seven brackets ranging from 10% to 39.6%.[14]

The message from all these reforms is unclear, except for a desire to protect the poor from taxation. The earned income tax credit, which grew rapidly during the 1990s, also greatly reduces the federal tax burden on the poor. At the same time, however, three of the other major U.S. taxes—the federal payroll tax, the state sales taxes, and the local property taxes—do not protect the poor from taxation.

A widely cited study of the U.S. tax system by Joseph Pechman and Benjamin Okner, last updated in 1984, estimated that the overall U.S. tax structure is mildly progressive at the lowest incomes and then roughly proportional over all remaining income levels. The low-end progressivity is due largely to the exemptions under the federal and state personal income taxes. (Other studies since 1984 have reached approximately the same conclusion.) The U.S. tax system does not redistribute much purchasing power in and of itself.[15]

REFLECTIONS ON THE HAIG–SIMONS CRITERION IN PRACTICE: THE FEDERAL PERSONAL INCOME TAX

Despite its appeal to many public sector economists, the Haig–Simons income measure has not fared well in the United States. Only the federal and state personal income taxes pay so much as lip service to the Haig–Simons criterion. State governments rely heavily on sales taxes and local governments rely primarily on the property taxes, neither of which is valid according to the Haig–Simons criterion.

State sales taxes may appear to be consistent with the view that consumption is the ideal tax base. In practice, however, sales taxes are far removed from an ideal consumption tax. Sales taxes often exclude broad classes of expenditures from taxation, they usually tax all included items at one fixed rate, and they are levied on businesses. What expenditures tax proponents have in mind is a tax levied on individuals exactly as the federal income tax is, except that it would exclude saving from the tax base. A graduated rate schedule could easily be applied to individual expenditures, removing the stigma from sales taxes that they may be regressive.[16]

The federal personal income tax is the single largest tax in the United States. Of all the broad-based taxes, it comes closest to the Haig–Simons income measure as its tax base, but not really all that close. Recall that, according to the Haig–Simons criterion, the federal income tax base should include personal income and capital gains on assets held from the beginning of the tax year, without regard to the sources or uses of income. The only permitted deductions from Haig–Simons income are legitimate business expenses, i.e., expenses required to earn the income. The actual tax base falls far short of the Haig–Simons ideal, both the personal income and capital gains components.

Personal Income

Taxable income is only about half of personal income. The main discrepancies between taxable and personal incomes fall into three categories: exemptions, exclusions, and deductions.

An *exemption* is an income that is recognized as taxable income by the Internal Revenue Service (IRS) but is simply not taxed. The main example is the personal exemption given to the taxpayer and all the taxpayer's dependents. The exemption was $3950 per person in 2014, and it is adjusted each year for increases in the consumer price index (CPI).

Exclusions are sources of income that are counted as personal income by the U.S. Department of Commerce in the National Income and Product.

Accounts are not counted as taxable income by the IRS. The principal exclusions are employee fringe benefits (primarily employer contributions to pension plans (along with the accrued interest on the investments under these plans), health insurance, and life insurance); interest income on IRAs and Roth IRAs, which are earmarked for retirement income; many federal, state, and local transfer

14. Tax Foundation, U.S. Federal Individual Income Tax Rates History, 1862–2013 (Nominal and Inflation Adjusted Brackets), http://taxfoundation.org/article/us-federal-individual-income-tax-rates-history-1913-2013-nominal-and-inflation-adjusted-brackets.

15. Okner and Pechman (1974). We will take a closer look at tax-incidence studies in Chapter 17.

16. This stigma may be more myth than reality. See the discussion in Chapter 17 on the incidence (burden) of the sales tax.

payments; imputed rental income on owner-occupied homes and imputed income on farm produce consumed on the farm; and interest received on state and local bonds, commonly referred to as "municipals."

Deductions are not sources of income at all, but rather certain expenditures that can be deducted from personal income in computing taxable income. The most important itemized deductions are extraordinary medical payments and other uninsured losses, state and local income and property taxes (taxpayers can elect to deduct state sales taxes instead of state income taxes), interest payments on mortgages for the principal residence, contributions to charities and other nonprofit organizations, and business expenses. Taxpayers can elect to take a "standard deduction" (equal to $12,400 in 2014 for married taxpayers filing jointly) instead of itemizing deductions.

The various exemptions, exclusions, and deductions exist because Congress and the administration cannot avoid the temptation to use the income tax to pursue other social ends, such as protecting low-income families and individuals from taxation, promoting homeownership, helping people save for their retirement years, subsidizing state and local governments, encouraging charitable giving, and so forth. These may all be worthy goals, but they come at a cost. The exemptions, exclusions, and deductions can undermine the horizontal and vertical equities of the tax. They also introduce inefficient distortions into the income tax. (We will return to these points below.)

Capital Gains

The capital gains portion of the tax base is also far from the Haig–Simons ideal. Recall that capital gains should be taxed as they accrue, and at the same rate applied to personal income. Also, capital losses should be fully offset against other income because they represent equal dollar decreases in purchasing power. Finally, the tax base should reflect increases in real purchasing power only. Increases in income arising solely from inflation should not be taxed. If nominal income is used as the tax base, then at least all sources of income should be treated equally with respect to the effects of inflation on purchasing power.

Capital gains taxation is deficient on all counts. Capital gains are taxed on a realized basis (that is, only when an asset is sold) and then at a separate rate from the rates applied to personal income (for assets held for more than 1 year), a rate that is often far below the rates applied to the higher-income brackets (20% in 2014); in effect, part of the realized gains is excluded from the tax base. The ability to offset losses against income is mostly limited to offsets against capital gains. Finally, the tax is levied on nominal capital gains, with no adjustment for the effects of inflation on purchasing power. As a result, capital and wage income are treated very differently in times of inflation.

Equity judgments about the income tax would be easiest if Haig–Simons income were the tax base (except for business expenses, which we will ignore from now on). Since Haig–Simons income is assumed to be the surrogate measure of utility, the tax payments themselves would be the appropriate basis for judgment. Horizontal equity would be satisfied if two persons with the same Haig–Simons income paid the same tax. Similarly, vertical equity would be appropriately measured by the difference in taxes paid by people with different amounts of Haig–Simons income.

Given the exemptions, exclusions, and deductions, however, the tax payments are no longer accurate measures of either horizontal or vertical equity. The problem is that markets react to any differences from the ideal tax base, and the market reactions have to be factored into any assessment of horizontal and vertical equities. They are sources of gains and losses to the taxpayers that matter every bit as much as the tax payments themselves. Tax burdens, not tax payments, determine the equity of the tax, and Feldstein's versions of horizontal and vertical equities defined in terms of utility are the only appropriate basis of judgment. For example, the proper statement of horizontal equity is that two persons with the same utility before tax should have the same utility after tax. That is, they should bear the same tax burden, the same loss of utility.

Unfortunately, judgments based on before and after utility comparisons can be problematic and are likely to confuse the general public. People tend to see the individual incomes and the tax payments, not the additional market-induced gains and losses. As a result, the exemptions, exclusions, and deductions are branded pejoratively as "tax loopholes." People see two taxpayers with the same Haig–Simons income who do not pay the same tax. Even worse, they see higher-income people paying less tax than lower-income people, an apparent equity reversal. The public's sense of horizontal and vertical equities is offended.

The Taxation of Personal Income: The Tax Loopholes

Not all tax loopholes in the personal income portion of the tax base are equally distasteful. The personal exemptions, for example, do not generate much complaint from the public. Protecting low-income people from taxation is an accepted goal in a nation that has declared a war on poverty. One of the often-stated criticisms of sales taxes, property taxes, and the payroll tax is that they do not offer such protection. Even proponents of a single, flat-rate

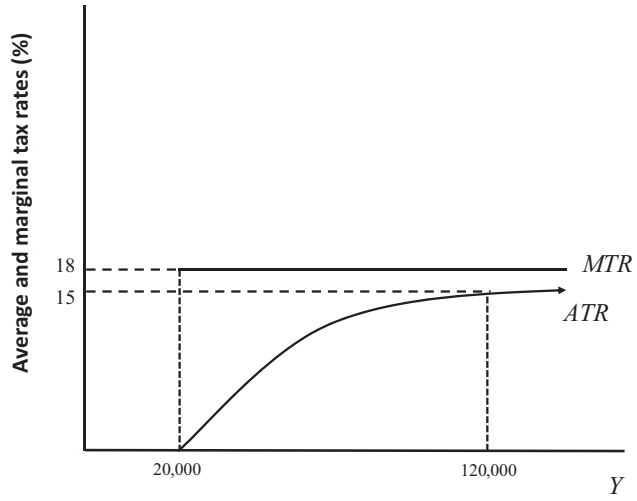

FIGURE 11.2

income tax favor including a personal exemption for the taxpayers and their dependents.

Exemptions are a simple way to ensure that a tax is progressive, if progressivity is desired. Fig. 11.2 illustrates the case of a taxpayer with three other dependents. It assumes a flat-rate tax of 18% on all income beyond a personal exemption of $5,000, or $20,000 for a family of four ($20,000 was approximately equal to the poverty line for a family of four in 2010). The vertical axis pictures the marginal and average tax rates. The marginal rates are 0 up to $20,000 and 18% thereafter. The average rates are also 0 up to $20,000 but then rise steadily beyond $20,000 as tax payments begin, approaching 18% asymptotically. (For example, at an income of $120,000, the tax is $18,000, and the average tax rate is 18/120 = 15%). The tax is progressive by the usual average tax rate measure.

The exclusions and deductions are far more contentious "loopholes," as perhaps they should be. They violate the pattern of vertical equity implicit in the tax structure, and they generate market and other forms of inefficiency. They may not be a source of horizontal inequity, however, despite the common perception that they are. The relationship between tax loopholes and horizontal equity is a particularly subtle issue that illustrates the importance of the market's reaction to the loopholes.

Tax Loopholes, Tax Capitalization, and Horizontal Equity

Consider the three large tax breaks to homeowners relative to those who rent an apartment: the exclusion of imputed rent on the home, the deduction for the interest payments on the mortgage, and the deduction for the local property taxes on the home. Fig. 11.3 illustrates the market's reaction to the tax break. The left-hand panel depicts the market for owner-occupied homes purchased by people within a certain income range (housing markets segment by income.) The right-hand panel depicts the market for rental apartments purchased by people within the same income range. The apartments are assumed to provide the same housing services as the owner-occupied homes, and the people in these markets are assumed to have identical tastes.

The equilibrium before these three tax breaks were introduced into the federal tax is given by the intersection of D^0 and S in each market. Since the housing services are identical, the prices are the same in each market, P_H^0 and P_A^0. (P_H^0 is the annualized price of a home, the implicit rental value). The people are indifferent to owning or renting.

The introduction of the three tax breaks makes the owner-occupied homes more attractive. Demand shifts up

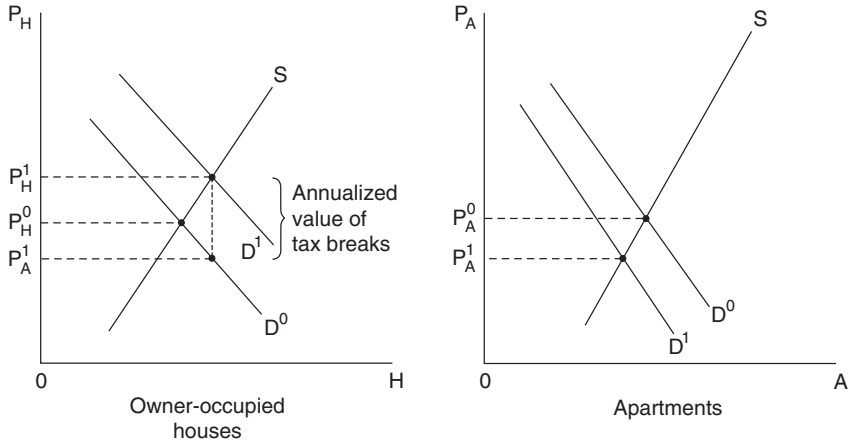

FIGURE 11.3

in the owner-occupied market and down in the apartment market, driving the (annualized) price of the homes up and the rentals on the apartments down. The new equilibrium occurs at the intersection of D^1 and S in each market, with the new equilibrium prices, P_H^1 and P_A^1.

At the new equilibrium, the difference in prices $P_H^1 - P_A^1$ must equal the annualized value of the three tax breaks to the homeowner. The market is said to *capitalize* the tax breaks into the relative prices of the two forms of housing. The implication of the market capitalization is that once the new equilibrium has been reached, the people in this income range are once again indifferent to owning or renting. If they choose to buy a house, the higher price minus the value of the tax breaks equals the rent they would have to pay for the apartment, P_A^1. This has to be the case, since the housing services are the same for the homes and the apartments, the people have identical tastes, and they are free to purchase a home or rent. Indifference to owning or renting is the only possible long-run equilibrium, regardless of the tax system.

This example illustrates the principle that any two persons in these markets, who had equal utility before the tax breaks were introduced, must have equal utility once the market returns to equilibrium in response to the tax breaks. The tax breaks do not violate horizontal equity in the new equilibrium. The homeowners get the tax breaks but no gain in utility relative to the renters. The same analysis applies to any tax loophole and for the same reason: The value of the loophole is eventually fully capitalized by the market system.

Feldstein summarized this fundamental principle of *tax design* as follows[17]:

> *"Once the market system establishes a long-run equilibrium in response to a given tax system, the tax system per se cannot be a source of horizontal inequity, where horizontal equity is defined in terms of burden or utility."*

A corollary to this fundamental principle of tax design is an equally fundamental principle of *tax reform:*

> *"Any reform of an existing tax code will create horizontal inequities through unanticipated gains and losses, and will continue to do so until a new long-run equilibrium obtains in the market place."*

Continuing with the housing example, suppose the three loopholes favoring homeownership were suddenly removed for promoting horizontal equity. Assuming the long-run equilibrium had been achieved, current homeowners surely lose, but not necessarily because they lose a tax advantage that had been unfairly given them, as the reformers intend. Rather, some of them will lose because they never received any gain in the first place at the higher prices they paid for their homes. These pure losses are an unavoidable consequence of any tax reform that removes the "loopholes."

A final point is that determining who gains from the three tax breaks is difficult once the market has reached its new equilibrium and the homes have changed hands a few times. Tax loopholes can even be capitalized in anticipation of the loopholes, before they become part of the law.[18] In conclusion, simply looking at tax payments gives a very misleading picture of horizontal equity when the tax contains various exclusions and deductions from Haig–Simons income.

Tax Loopholes, Vertical Equity, and Inefficiency

Although tax loopholes may not be a source of horizontal inequity, tax reformers can still make a good case for removing them. They are likely to give rise to vertical inequities, and they lead to various kinds of inefficiencies. Therefore, the gains to vertical equity and efficiency from removing the loopholes may exceed any temporary horizontal inequities plus the lost benefits associated with whatever social goals the loopholes are trying to promote. This is especially so if there are more effective ways of promoting the social goals.

The housing example above illustrates the possibility of vertical inequity. Both homeowners and apartment renters gain equally from the three tax breaks. Their

17. See Feldstein (1976), pp. 94–97.

18. For further analysis of the owner-occupied tax breaks that include supply adjustments, see White and White (1977). Boris Bittker tells an amusing anecdote illustrating the principle of capitalization. It concerns an eager, young law student who searches in vain for the beneficiary of a tax-sheltered apartment building in his hometown known as Rainbow Gardens. The tax shelter had been in existence since the inception of federal income taxation under the Revenue Act of 1913. Rainbow Gardens was for sale, but the law student quickly surmised that at the asking price he would only realize a normal return on his investment. He also learned that the current owners were selling because they, too, were only able to earn a normal return despite the existence of the tax shelter. The same had been true of the previous owners, and the ones before them, and so on. Alas, they all paid too much to realize an economic profit from the tax shelter. The persistent student was able to trace the line of ownership all the way back to R. E. Greison, who had purchased Rainbow Gardens in 1896. Greison possessed a remarkable foresight. In 1896, he was clerking for a U.S. Supreme Court justice when the Court ruled that a federal income tax was unconstitutional. Greison nonetheless correctly predicted that the Court's decision would eventually be overturned by a constitutional amendment (the 16th), and further that the income tax law, when drafted, would tax shelter apartment buildings. Based on these predictions, Greison bought Rainbow Gardens. Sad to say, the capitalization of the tax shelter predated Greison. His epitaph read: "Sacred to the memory of R. E. Greison, who learned that before every early bird, there is an earlier bird." See Bittker (1975).

annual costs fall from $P_A^0 (= P_H^0)$ to P_A^1 because of the tax breaks, whether they own or rent. But, as noted above, housing markets segment by income. It is possible, therefore, that the decrease in their housing costs is greater than the decrease in the housing costs of other lower-income people. If so, the larger break to the higher-income people is likely to offend people's sense of vertical equity.

Exclusions and deductions always generate this kind of vertical inequity under an income tax with graduated rates. Since the exclusion or deduction is taken off the tax base, it reduces the taxpayer's liability by t cents per dollar of exclusion or deduction, where t is the taxpayer's marginal tax rate. Under a graduated tax, the value of the tax savings rises with income. For example, every dollar given to charity that can be deducted from taxable income costs the taxpayer only \$ $(1 - t)$. The richer the taxpayers, the lower their costs of contributing to their favorite charity, church, or school. This is why economists tend to favor tax credits over exclusions or deductions if the tax system is to be used to encourage certain activities. A 10% tax credit is taken directly against the tax liability, after the tax has been computed, so that it is 10% for all taxpayers.

The housing example also illustrates the market inefficiencies of tax loopholes. As we will learn in Chapter 13, anything that drives a market away from its normal supply and demand equilibrium generates a deadweight efficiency loss. In the example above, the natural equilibria in the two markets are at the intersection of S and D^0 in the two markets, without the tax breaks. The tax breaks lead to too many homes and too few apartments, with resulting dead-weight efficiency losses in the market for houses.[19] Tax loopholes inevitably lead markets away from their natural equilibria as the markets capitalize the loopholes. Thus, they necessarily generate efficiency losses (unless supply or demand is perfectly inelastic, which is rare in the long run).

Tax loopholes lead to other inefficiencies as well. Consider the exclusion for interest received on state and local bonds, the municipals. The exclusion acts as a subsidy to the lower-level governments, equal to the reduction in debt service made possible by the municipals' tax-free status, but this is a particularly inefficient form of subsidy from the federal government's point of view. Suppose a state government can offer an interest rate of 8% rather than 10% because of the exclusion, a savings of \$20 of interest income on each \$1000 bond. The problem is that only investors who can save more than \$20 in taxes will purchase the municipals. For example, at the assumed interest rates, a person in the 28% tax bracket can earn interest of \$80 net of tax on the municipal at 8% but only \$72 net of tax on the taxable bond at 10%. Therefore, the U.S. Treasury loses more than \$1 in tax revenue for every \$1 of interest subsidy received by a state or locality, \$28 of lost revenue for a \$20 subsidy in this example. In contrast, a direct federal subsidy (grant-in-aid) for capital expenditures would give \$1 of subsidy for every \$1 of tax revenue collected, a more efficient subsidy from the federal government's viewpoint. Also, the direct subsidy avoids the deadweight loss inefficiencies in the bond market as the tax break to the municipals is capitalized into a lower interest rate relative to the taxable bonds.

Another source of efficiency gain from removing the loopholes is that the tax becomes much simpler, which saves on administrative and compliance costs. In addition, a broader tax base means that the same tax revenues can be collected with lower tax rates, which sharply reduces the size of the deadweight loss in the marketplace. We will see in Chapter 13 that the deadweight efficiency loss from a tax varies directly with the square of the tax rate.

Would the gains to vertical equity and efficiency from removing the exclusions and deductions more than offset any temporary horizontal inequities that may arise and the social benefits of the loopholes? This remains an open question, but many economists favor removing most of the exclusions and deductions. The economists in the Treasury Department during the Reagan administration put forth such a plan in their proposal for TRA86. The administration's proposal called for almost a textbook version of an income tax based on Haig−Simons income, with little more than the personal exemptions and legitimate business expenses as reductions to the tax base. The administration's proposal could not stand up to the special interest groups favoring the loopholes, however, and all the major exclusions and deductions were retained. No major tax base reform proposals have received a serious hearing in Congress since 1986.

The Taxation of Capital Gains: Inflation Bias and Realization

Two long-standing issues in the taxation of capital gains are that capital gains (and other sources of income from capital) are not protected from inflation in computing taxable income and that capital gains are taxed on a realized rather than an accrued basis. The failure to index income from capital for inflation can lead to a huge bias against income from capital in an inflationary economy, a troubling situation for the United States, given its relatively low rates of

19. We will see in Chapter 13 that efficiency losses arise only in markets that are distorted by features such as taxes or subsidies and market power. In this example, the competitive apartment market is not distorted and thus not a source of efficiency loss.

saving and investment. Taxing capital gains on a realized rather than an accrued basis generates further sources of inefficiencies and vertical inequities. The next two sections consider these issues.

THE INFLATIONARY BIAS AGAINST INCOME FROM CAPITAL[20]

The U.S. Tax Codes were written for a noninflationary economy. This, in itself, generates an extra tax burden on income from capital relative to wage income, simply because inflation causes nominal asset income to grow more rapidly than nominal wage income. Consequently, equal growth in nominal income from these two sources reflects unequal growth in real purchasing power, so that equal taxation implies unequal treatment, in violation of horizontal equity.[21] The inequity is compounded by the graduated rate schedule, because the artificially expanded tax base of asset income may be taxed at higher rates.

To see how the differential inflation effect arises, assume that an economy has been experiencing inflation since time $t = 0$. Define the accumulated inflation to time t as

$$I(t) = \exp \int_0^t i(s)\mathrm{d}s \tag{11.12}$$

where $i(t) =$ the instantaneous rate of inflation at time t. Assume further that inflation is fully anticipated so that $i(t)$ represents both the actual and expected rates of inflation. If $W(t)$ represents wage income at time t without inflation, then

$$W'(t) = W(t) \cdot I(t) \tag{11.13}$$

is wage income with inflation.

Let $Y(t)$ represent income from capital in the absence of inflation:

$$Y(t) = r(t) \cdot V(t) \tag{11.14}$$

where $r(t) =$ the real rate of return and $V(t) =$ the value of an asset without inflation. The basis arises because expected inflations affects income from capital in two ways. It increases both the value of assets *and* the rate of return on assets. Let

$$V'(t) = V(t) \cdot I(t) \tag{11.15}$$

represent the value of assets with inflation, and

$$n(t) = r(t) + i(t) \tag{11.16}$$

represent the nominal rate of return. Hence

$$Y'(t) = n(t) \cdot V'(t) = [r(t) + i(t)]V(t) \cdot I(t) \tag{11.17}$$

where $Y'(t) =$ income from capital with inflation. Dividing Eqn (11.17) by Eqn (11.13), using Eqns (11.14) and (11.15), and rearranging terms yields

$$\begin{aligned} \frac{Y'(t)}{W'(t)} &= \frac{(r(t) + i(t)) \cdot V(t) \cdot I(t)}{W(t) \cdot I(t)} \\ &= \frac{Y(t)}{W(t)} + \frac{i(t)V'(t)}{W'(t)} > \frac{Y(t)}{W(t)} \end{aligned} \tag{11.18}$$

Therefore, capital income grows more rapidly than wage income simply because of the inflation factor. If the tax base is nominal income, capital income is overly taxed. By inspection of the right-hand side (RHS) of Eqn (11.18), the inflationary bias can be removed by subtracting from nominal asset income the expected rate of inflation times the value of assets with inflation, $i(t) \cdot V'(t)$, before applying the tax rates.

The inflation adjustment should be applied to all sources of capital income, but the nature of the adjustment varies depending upon the particular form of the asset. For example, if the income derives from an interest-bearing asset, the taxable income should include only the proportion of the interest resulting from the real rate of return. That is, if $Y'(t) = (r(t) + i(t)) \cdot V'(t)$, then

$$Y'(t) - i(t) \cdot V'(t) = r(t)V'(t) \tag{11.19}$$

But,

$$r(t)V'(t) = \frac{r(t)}{n(t)} \cdot n(t) \cdot V'(t) = \frac{r(t)}{n(t)} \cdot Y'(t) \tag{11.20}$$

Thus, the taxpayer would report actual interest payments times the ratio of the real to the nominal rate of return.

For a straight capital gain without interest payments, the taxpayer would increase the purchase price by the accumulated inflation factor before subtracting it from the current value to compute the capital gain. For these assets,

$$Y'(t) = \mathrm{CV} - \mathrm{PV} \tag{11.21}$$

where

CV = current value, inclusive of inflation.
PV = original purchase value.
Adjusting $Y'(t)$ yields

$$Y'(t)_{\text{adjusted}} = (\mathrm{CV} - \mathrm{PV}) - (\mathrm{CV} - \mathrm{PV})_{\text{inflation}} \tag{11.22}$$

$$Y'(t)_{\text{adjusted}} = (\mathrm{CV} - \mathrm{PV}) - (\mathrm{PV} \cdot I(t) - \mathrm{PV}) \tag{11.23}$$

20. The analysis in this section follows Diamond (1975), pp. 228–230.
21. Vertical equity is also necessarily violated by the inflation bias, presumably in an anti-rich, pro-poor direction. Horizontal equity may not be violated if investments in physical and human capital are perfect substitutes. In that case, the inflation bias would drive investment toward human capital, lowering wages and raising the return to physical capital until the difference just equaled the value of the inflation bias against physical capital.

$$Y'(t)_{\text{adjusted}} = \text{CV} - \text{PV}\cdot I(t) \qquad (11.24)$$

Finally, money holdings would receive a credit equal to $i(t)\cdot V'(t)$, since there is no nominal return from which to subtract this adjustment factor. Diamond recommends ignoring this adjustment on the grounds that the bookkeeping for cash assets would be especially difficult and that the liquidity services from holding cash are untaxed anyway.[22]

These inflation adjustments are correct as given only if inflation is always fully anticipated and all income inflates at the same rate over time. In actuality, of course, neither of these is true, and it is not clear what should be done to correct the discrepancies. As a practical matter, governments would surely have to use actual rather than expected rates of inflation for any adjustment. Economic research has not even been able to determine how inflationary expectations are formed. Yet nominal interest rates and capital values almost certainly adjust to some degree for anticipated inflation. Thus, it is probably more accurate to adjust capital income by some long-run smoothed inflation index than to make no adjustment at all. People who anticipate inflation incorrectly will either make windfall gains or losses relative to the theoretical ideal, but this is unavoidable.

The practical question remains as to which long-run series to use, since assets and other sources of income inflate at different rates. A broad series such as the CPI is probably a reasonable choice for practical purposes, although again some people will receive (suffer) windfall gains (losses) in purchasing power relative to the ideal.

TAXING REALIZED GAINS: AUERBACH'S RETROSPECTIVE TAXATION PROPOSAL

Economists have long understood the equity and efficiency problems caused by taxing capital gains on a realized basis rather than an accrued basis. The equity problem is that deferring taxes on the capital gains until they are realized places the government in the position of offering interest-free loans each year to the asset holders. The loans equal the amount of the tax liability that would have been paid on an accrued basis. Since assets that generate capital gains, such as common stocks, are disproportionately held by the rich, the pattern of loans is likely to violate the public's sense of vertical equity. The efficiency problem is that taxing capital gains on a realized basis alters the pattern of buying and selling of assets that would occur if the gains were taxed properly on an accrued basis. By taxing on a realized basis, investors have an incentive to "lock-in" the gains on successful assets (choose not to sell) to defer the tax payment and to sell unsuccessful assets early to deduct the losses against other sources of income. A related inefficiency is the incentive to take income as capital gains to defer the tax, an example being executives who take stock options in lieu of salary.

The inequities and inefficiencies notwithstanding, no one has seriously proposed taxing capital gains on an accrued basis. The difficulty comes with assessing the accrued gains on real assets that are infrequently traded. How much capital gain accrued last year on the house that has not been on the market since 1980 or the painting that has been hanging in the den since 1985? Tax authorities have no good way of estimating the gains (or losses) for these assets. Even if they could evaluate the accrued gains, people whose wealth consisted primarily of real assets may have to sell some of their assets to pay the accrued tax liability. The public would tend to view this as unfair. (An analogous situation is the elderly couple that is forced to sell their house they have lived in for 50 years because they can no longer pay the local property taxes out of their retirement pension.) For all these reasons, capital gains will almost certainly continue to be taxed on a realized basis.

In 1991, Alan Auerbach achieved a substantial breakthrough in solving the problems of taxing gains on a realized basis. He proposed a tax reform that avoids the lock-in and early sales effects by leaving investors always indifferent between: (1) holding an asset for one more period and (2) selling the asset and investing the after-tax proceeds in a risk-free asset for one period. His proposal also protects the government from making interest-free loans, at least on an expected value basis. The beauty of the Auerbach proposal is its practicality. It continues to tax capital gains on a realized basis and makes use of data that are readily available at the time the asset is sold.

A Two-Period Example

The following simple two-period example provides the intuition for the nature of the realization problem and how Auerbach proposes to overcome it. Consider two options for investing $1 at the start of the first period.

Option 1*:* Hold the asset for one period, realize the gain at the end of the period, and invest the after-tax proceeds in a risk-free asset during the second period.

Option 2: Hold the asset for two periods and then realize the capital gain over the two periods.

Assume:

g = the capital gain during the first period
i = the one-period return on the risk-free asset
r = the (uncertain) capital gain during the second period
t = the income tax rate.

22. See Diamond (1975), p. 232.

Option *1: Sell and Invest Risk-Free Asset*

The value of the asset at the end of the first period is $(1+g)$. The realized gain g is taxed at rate t, leaving net of tax proceeds of $[1+g(1-t)]$ to be invested in the risk-free asset during the second period. The proceeds grow at rate i and the interest is taxed at rate t. Therefore, the net-of-tax value of the asset at the end of period 2 is

$$[1+g(1-t)][1+i(1-t)]$$

For comparison with option 2, rewrite the net-of-tax value as

$$[(1+g)-tg][(1+i)-it] = [(1+g)(1+i)] - t[(1+g)i + g(1+i(1-t))]$$

The first bracketed term is the gross-of-tax value and the second the tax liability.

Option *2: Hold for Two Periods*

The value of the asset at the end of period 2 is $(1+g)(1+r)$, and a tax is paid on the capital gain, leaving a net-of-tax value at the end of period 2 equal to:

$$[(1+g)(1+r)] - t[(1+g)(1+r)-1]$$
$$= [(1+g)(1+r)] - t[(1+g)r+g]$$

On comparing the two outcomes, note that r is an uncertain return at the end of period 1. Assume that the certainty equivalent of r is i. That is, investors are indifferent between investing at the uncertain return r or the certain return i.[23] Under this assumption, the certainty equivalent net-of-tax value of option 2 is

$$(1+g)(1+r) - t[(1+g)i+g]$$

Thus, option 2 is more valuable by the amount $(tg)i(1-t)$, equal to the after-tax interest on the portion of the accrued tax liability that is avoided by taxing the capital gain on a realized basis. The tax savings can be thought of as an interest-free loan by the government (tg) made at the end of period 1. The taxpayer invests the risk-free loan at rate i during the second period, pays a tax on the interest at rate t, and pays back the principal on the loan, for a net gain of $(tg)i(1-t)$, the after-tax interest on the loan.

The value to the asset holder of taxing on a realized basis follows the same pattern for any number of periods. The value equals the net-of-tax interest on the current value of the taxes that would have been collected each year if capital gains were taxed on an accrued basis. (Note that the tax is paid once, when the asset is sold. The deferred tax liabilities, the "loans," accumulate at untaxed interest until the sales date.)

23. $i = E[r]$ under risk neutrality.

The Vickrey Proposal

In 1939, William Vickrey proposed the following tax on capital gains to remove the interest-free loan advantage from taxing on a realized basis: Tax the gain in the final period on a realized basis, and add to the tax the interest on the current value of accrued tax liabilities to date, with the interest being tax deductible (Vickrey, 1939). The combined tax plus interest payment would make asset holders indifferent at any given time between holding the asset for one more period or selling the asset and investing the after-tax proceeds in a risk-free asset.

Under the Vickrey scheme, the instantaneous *increase* in the tax at time s if the asset is held one more period is, in general,

$$\dot{T}_s = i(1-t)T_s + tr_sA_s \qquad (11.25)$$

where

T_s = the current value of the accumulated deferred tax liabilities to date at time s.

r_s = the gain in period $s+1$.

A_s = the current value of the asset at time s.

In terms of the two-period example above, s is the end of period 1, $T_s = gt$, and $tr_sA_s = tr(1+g)$.[24]

The problem with Vickrey's scheme is that it is as impractical as taxing on an accrued basis. It requires knowing the entire pattern of accrued tax liabilities to the time of sale, which is the same as knowing the entire pattern of gains. The current value of the total accrued taxes due on an asset held for 10 years is quite different if all the gains came in the first year, or in the last year, or evenly over time. In other words, the data requirements are the same as they would be under an accrued tax, data that would be unavailable for infrequently sold real assets.

The Auerbach Proposal

Auerbach (1991) proposed a variation of the Vickrey scheme that is practicable for all assets. It is based on the certainty equivalence operator, $V(\ )$, which gives the value that an investor would require, with certainty, to be indifferent to an uncertain return that is the argument of the function V. The idea is that investors make their portfolio choices prospectively. They are indifferent to holding an uncertain asset for one more period if the certainty equivalence of the after-tax return on the asset is equal to the risk-free after-tax return. In terms of the operator V, indifference requires that at time s

$$V(\dot{A}_s - \dot{T}_s)/(A_s - T_s) = i(1-t) \qquad (11.26)$$

24. The part of the realized tax liability, gt in the two-period example, is the tax on the first-period gain. It is not part of the tax increase.

where $\dot{A}_s$ is the uncertain next period return on the asset, and the other terms are as defined above. Multiplying both sides of Eqn (11.26) by $(A_s - T_s)$ yields

$$V(\dot{A}_s - \dot{T}_s) = i(1-t)[A_s - T_s] \tag{11.27}$$

$V(\ )$ is a linear operator. Therefore,

$$V(\dot{A}_s) - V(\dot{T}_s) = iA_s - i(1-t)T_s - itA_s] \tag{11.28}$$

But $V(\dot{A}_s) = iA_s$, the certainty equivalent next period return on the asset. Thus, indifference to holding or selling requires that

$$V(\dot{T}_s) = i(1-t)T_s + itA_s \tag{11.29}$$

Auerbach's proposal is Vickrey's proposal from an ex ante rather than an ex post perspective. $V(\dot{T}_s)$, the ex ante certainty equivalence of the increase in the taxes, is the net-of-tax risk-free interest on the deferred tax liabilities plus the tax on the certainty equivalent return for the next period.

Auerbach proves that the required $V(\dot{T}_s)$ is achieved if and only if the accumulated tax liability upon realization, T_s, is

$$T_s = (1 - e^{-its})A_s \tag{11.30}$$

Note that T_s depends only on current data at the time the asset is sold: the risk-free market interest rate, i; the number of periods that the asset has been held, s; the marginal tax rate, t; and the current value of the asset, A_s. The taxpayer could easily determine the tax liability by looking it up in a table. Note, also, that $T_s = 0$ when $s = 0$ (an asset bought and sold immediately yields no income and incurs no tax); and $T_s \to A_s$ as $s \to \infty$ (the accumulated deferred tax approaches the value of the asset as the holding period extends into the future without limit).

We will demonstrate the sufficient conditions and leave the necessary conditions to the interested reader.[25] Suppose $T_s = (1 - e^{its})A_s$. Then, the instantaneous increase in taxes from holding one more period is

$$\dot{T}_s = (1 - e^{-its})\dot{A}_s + it\, e^{-its} A_s \tag{11.31}$$

Add and subtract A_s to the RHS and multiply and divide the first term by A_s:

$$\dot{T}_s = (1 - e^{-its})(\dot{A}_s/A_s)A_s - (1 - e^{-its})it\, A_s + it\, A_s \tag{11.32}$$

$$\dot{T}_s = (1 - e^{-its})A_s[(\dot{A}_s/A_s) - it] + it\, A_s \tag{11.33}$$

But, $\dot{A}_s/A_s = i + e$, where e is a random variable with mean zero. Thus,

$$\dot{T}_s = (1 - e^{-its})A_s[(i + e - it)] + it\, A_s \tag{11.34}$$

$$\dot{T}_s = (1 - e^{-its})A_s[(e + i(1-t)] + it\, A_s \tag{11.35}$$

But $V(e) = 0$, by definition of the certainty equivalence operator. Therefore,

$$V\dot{T}_s = (1 - e^{-its})A_s[i(1-t)] + it\, A_s \tag{11.36}$$

or

$$V\dot{T}_s = T_s[i(1-t)]) + it\, A_s \tag{11.37}$$

as required for investor indifference for holding the asset one more period or realizing and investing in the risk-free asset.[26]

The only caveat to Auerbach's proposal is that the tax is essentially a prospective tax because it is based on the certainty equivalence of the next period return rather than the actual return. Many proponents of income taxation tend to believe that the fair way to tax is on the ex post actual returns and not the ex ante expected returns. From the ex post perspective, exceptionally good assets are undertaxed and exceptionally poor assets are overtaxed under Auerbach's proposal. Nonetheless, investors do base their decisions on prospective returns, so that Auerbach's proposal does avoid the lock-in effect. Whether it is entirely fair or not depends on the ex post versus ex ante point of view, and this is largely a matter of taste.

Economists who favor expenditures taxes based on lifetime utility arguments tend to be indifferent between taxing on an ex post or ex ante basis. For example, they are indifferent between taxing the value of a house when it is purchased or the stream of housing services as they accrue, because the purchase price of the house equals the *expected* present value of the stream of housing services. The IRS would tax the value of the house when purchased under an expenditures tax because it is the only practical alternative. Whatever one's view of its equity implications, Auerbach's proposal for taxing-realized capital gains must be considered a landmark in the theory of tax design for having solved the capital gains lock-in problem in a practical manner.

Capital Gains Taxation: A Postscript

Congress has never protected income from capital from inflation nor even remotely considered adopting Auerbach's tax scheme. Instead, it has favored either excluding a portion of "long-term" capital gains, the gains on assets held for more than 1 year, thereby effectively taxing the

25. The necessary conditions establish that Eqn (11.30) is the only possible T_s that is a function of only i, s, t, and A_s. See Auerbach (1991), pp. 172–173.

26. Auerbach also presents more complicated cases, such as the appropriate tax for indifference when there are both capital gains and dividends.

gains at a lower rate, or, as this is written, taxing capital gains at a single, relatively low rate. Favoring capital gains is done in the name of encouraging saving. It is also justified as a way of offsetting the inflationary bias against capital gains. At the same time, however, it has the effect of giving another tax break to high-income taxpayers in addition to the interest-free loans they receive from deferring the tax until realization.

THE TAXATION OF HUMAN CAPITAL

Louis Kaplow (1996) has taken a provocative position regarding the appropriate taxation of wage income under an ideal income tax if one views wages as the returns to a person's stock of human capital. His point is simply that physical and human capital should be treated identically under an ideal income tax. If this were done, however, it would lead to a sharp increase in the share of taxes collected from income received by labor.

Wages are not treated as returns to human capital in the standard Haig–Simons version of the ideal income tax presented earlier in this chapter. Instead wages are viewed as arising completely and concurrently with the supply of labor and are therefore taxed in full as they are realized each year, exactly as they would be taxed under an ideal wage or payroll tax. Viewing wages as returns to human capital would lead to very different tax treatment.

Compare, for example, the decision to save and the decision to invest in human capital. Under an ideal income tax, the saving is taxed immediately (not deducted from taxable income), and the returns to the saving are taxed as they *accrue*, whatever form they may take (e.g., interest, capital gains, a stream of returns from a real, depreciable asset). In the case of a real asset, the taxable returns are the gross returns less the depreciation on the asset each year, with the depreciation equal to the decline in the value of the asset. An investment in human capital is most directly equivalent to an investment in a real, depreciable asset. The initial investment costs should be taxed, that is, not deducted from taxable income. Also, the net returns from the investment—equal to the increase in the wages less the annual depreciation of the stock of human capital—should be taxed each year. Neither requirement is met under the standard income tax. Investment in the human capital may well be expensed, that is, deducted in full from taxable income, if it takes the form of lower wages received while participating in an on-the-job training program. Also, no deduction is allowed for the depreciation of a person's human capital. The wages are taxed in their entirety each year as they are realized. Notice that, from the human capital perspective, the full taxation of wages each year is completely wrong in the person's last year of work because the stock of human capital necessarily depreciates to zero in the last year, and always by an amount equal to the wages earned in that year. The final-year tax liability should be zero under an ideal accrued income tax.

Kaplow takes the taxation of human capital one step further by assuming that all wages can be thought of as a return to human capital. Under this view, the stock of human capital is a gift received at birth that should be subjected to two forms of taxation if treated symmetrically to physical capital.

First, the receipt of a gift of physical capital, or of any financial asset that is ultimately a claim against the earnings of physical capital, is treated as income under an ideal income tax and subject to full taxation. Therefore, the initial gift of human capital should be treated as income and subject to full taxation at birth. The value of the gift is the present value of a person's lifetime stream of wages less any expenses/investments incurred to generate the wage stream. In a world of perfect certainty, all future expenses/investments associated with the maintaining and increasing the stock of human capital would be known at birth, as would the entire stream of future wages arising from the human capital. The cash flow from the human capital would be lower in years in which future investments were made and higher in the noninvestment years. In other words, the initial gift of human capital at birth is its capacity to engage in certain kinds of investments in human capital throughout one's lifetime, along with the lifetime wages that result from the investments.

In addition, any accrued income (net of depreciation) earned by the physical capital gift in subsequent years is subject to taxation each year. Similarly, the stream of wages each year net of depreciation resulting from the gift of human capital should also be taxed under an ideal income tax. Given the usual pattern of depreciation of human capital, the present value of the depreciation is likely to be less than the present value of the wage stream because wages will far exceed depreciation except in the last working years. Thus, the annual stream of wages and depreciation represents a second source of taxable income.

To summarize, the appropriate tax base for human capital under an ideal, accrued income tax consists of: (1) the initial gift of human capital at birth, equal to the present value of lifetime wages less any lifetime expenses/investments and (2) the annual stream of wages less the depreciation of the stock of human capital. This tax treatment is equivalent to the ideal tax treatment of a gift of physical capital.

In fact, gifts of physical capital (or financial assets) are stepped up in basis when passed on to heirs, so that the initial value of the capital escapes taxation. This is not supposed to happen under an ideal income tax, but because it does happen one could argue for exempting the initial gift of human capital from taxation. If so, then the tax base for

human capital is just the annual stream of wages less depreciation of the human capital stock. Since the standard "ideal" income tax calls for full taxation of wages, it actually overtaxes wage income when it is viewed as a return to human capital.

Suppose the income tax were reformed to include all gifts of capital as income, as called for by an ideal income tax. Then, if human capital escapes taxation at birth, the Auerbach/Vickrey method of retrospective taxation could be employed to capture the escaped tax liability when the wages (returns to human capital) are realized. The taxes due on the wages each year would include tax-deductible interest since birth on the taxes that should have been collected on those wages at birth. The later in life that the wages occur, the higher the tax due on them, because the taxes due on them at birth have been receiving implicit interest tax free since then.

For example, the present value at birth of \$1 of wages received at time i equals $\frac{1}{(1+r^i)}$, where r is the annual gross-of-tax interest rate, assumed constant over time. Had a tax been collected on that wage at birth, the value of the human capital at time i would have been $(1-t)\left[\frac{(1+r_a)}{(1+r)}\right]^i$ where t is the tax rate and r_a is the after-tax rate of interest, both assumed constant over time. Therefore, the *current value* of the taxes that should have been collected at birth, increased by the after-tax interest rate since birth, is

$$\left[1-(1-t)\left[\frac{(1+r_a)}{(1+r)}\right]^i\right]$$

If the escaped taxes are to be collected retrospectively at rate t in period i, then the \$1 of wages has to be scaled by $\left[\frac{1}{t}-\frac{(1-t)}{t}\left[\frac{(1+r_a)}{(1+r)}\right]^i\right]$ to collect this portion of the tax due under the ideal income tax.

Kaplow presents some calculations to show that wages received in the last few working years would have to be increased by a factor of 2−3 to capture retrospectively the escaped taxes since birth. Such scaling of the wages would be equivalent to scaling the returns on tax-deferred pension instruments such as IRAs if the taxes were collected retrospectively when the returns were realized. Under an ideal accrued income tax, savings for retirement should not be deducted from income as IRAs are.

Taxing human capital at birth, or scaling wages later on in life to account for taxes that should have been paid, would undoubtedly lead to horrendous problems of evaluation and liquidity—people might not trust how the tax liabilities were calculated or be able to pay the taxes when they are due. The fact remains, however, that Kaplow's suggested treatment of human capital is the proper one under an ideal accrued income tax if wages are viewed as the returns to human capital.

The discussion so far has assumed perfect certainty. The taxation of uncertain income streams would be resolved as the uncertainty is resolved: Unexpected favorable (unfavorable) returns to human capital would increase (decrease) its value and the taxes due.

In conclusion, the only three ways that taxable income from human capital can arise are at birth, over time (the stream of wages less depreciation), and as uncertainties about future income streams are resolved. Proponents of the ideal taxation of physical capital should favor similar taxation of human capital if they view wages as a return to human capital. This point takes on special force given the widely cited estimate by James Davies and John Whalley that the stock of human capital in the United States is on the order of three times the stock of physical capital.[27] If gifts of human and physical capital were counted as income as they should be, then the share of tax revenues collected from labor income would rise substantially.

SUMMARY

This chapter has emphasized that the problem of designing equitable broad-based taxes is one of the more vexing in all of public sector economics. First-best theory offers two guidelines for tax design: the interpersonal equity conditions of social welfare maximization and the ability-to-pay principle. The interpersonal equity conditions are preferred by the mainstream theory, yet the ability-to-pay principle has won the day in terms of informing tax policy. Even so, ability-to-pay principles are subject to various interpretations. Furthermore, even if ability-to-pay principles can be agreed upon, it is extremely difficult to determine who has actually gained or lost from a given tax system and who will gain or lose from particular tax reforms.

A brief review of the U.S. federal personal income tax served to highlight these problems. The tax pays lip service to the ability-to-pay principle on paper, but there are many slips in application. The chapter considered a number of reforms that would make the tax conform more closely to traditional ability-to-pay principles, such as removing certain exclusions and deductions. But we were forced to admit that these reforms would not necessarily make the tax more equitable under a proper utility-based interpretation of these same principles, since reforms themselves generate inequities. Equity in taxation is as difficult to achieve as equity in any other context.

27. Davies and Whalley (1991). Kaplan is definitely not proposing that the personal income tax be reformed to treat wages as returns to human capital. To the contrary, he does not believe that the ability-to-pay perspective is a useful addition to tax theory. He prefers the modern social welfare function perspective on taxes, transfers, and distributive equity generally which, as discussed earlier in the chapter, is concerned much more with issues of the tax structure (vertical equity) than with precisely defining the tax base.

REFERENCES

Auerbach, A., March 1991. Retrospective capital gains taxation. American Economic Review 81 (1), 167–178.

Auerbach, A., June 2006. The choice between income and consumption taxes: a primer. NBER Working Paper No. 12307.

Bittker, B., December 1975. Tax shelters and tax capitalization or does the early bird get a free lunch. National Tax Journal 28 (4), 416–419.

Davies, J., Whalley, J., 1991. Taxes and capital formation: how important is human capital? In: Bernheim, B.D., Shoven, J. (Eds.), National Saving and Economic Performance. University of Chicago Press, Chicago.

Diamond, P., August 1975. Inflation and the comprehensive tax base. Journal of Public Economics 4 (3), 227–244.

Feldstein, M., July-August 1976. On the theory of tax reform. Journal of Public Economics 6 (1-2), 77–104 (International Seminar in Public Economics and Tax Theory).

Haig, R.M., 1921. The concept of income: economic and legal aspects. In: Haig, R.M. (Ed.), The Federal Income Tax. Columbia University Press, New York.

Kaldor, N., 1955. An Expenditure Tax. George Allen & Unwin, Ltd., London.

Kaplow, L., May 1996. On the divergence between 'Ideal' and conventional income-tax treatment of human capital. American Economic Review 86 (2), 347–352.

Mill, J.S., 1921. In: Ashley, W.J. (Ed.), Principles of Political Economy. Longmans, Green Co., Ltd, London.

Musgrave, R., 1959. The Theory of Public Finance. McGraw-Hill, New York.

Musgrave, R., June 1990. Horizontal equity, once more. National Tax Journal 43 (2), 113–122.

Okner, B., Pechman, J., 1974. Who Bears the Tax Burden? The Brookings Institution, Washington, D.C, p. 10.

Pechman, J. (Ed.), 1980. What Should Be Taxed? Income or Expenditure. The Brookings Institution, Washington, D.C.

Sadka, E., December 1976. On progressive taxation. American Economic Review 66 (5), 931–935.

Simons, H.C., 1938. Personal Income Taxation. University of Chicago Press, Chicago.

Smith, A., 1904. In: Cannan, E. (Ed.), The Wealth of Nations, vol. II. G. P. Putnam's Sons, New York.

Vickrey, W., June 1939. Averaging income for income tax purposes. Journal of Political Economy 47 (3), 379–397.

White, L.J., White, M.J., February 1977. The tax subsidy to owner-occupied housing—who benefits? Journal of Public Economics 7 (1), 111–126.

Young, H., April 1988. Distributive justice in taxation. Journal of Economic Theory 44 (2), 321–335.

Young, H., March 1990. Progressive taxation and equal sacrifice. American Economic Review 80 (1), 253–266.

Chapter 12

Introduction to Second-Best Analysis

Chapter Outline

First-best analysis offers us a complete, internally consistent normative theory of the public sector, yet the theory is far from satisfactory. It is often quite unrealistic and therefore unresponsive to the needs of policy makers. First-best models ignore a number of important real-world phenomena that the policy maker cannot ignore.

The strengths and weaknesses of first-best theory derive from a common source, that the only restrictions on a first-best policy environment are the two sets of restrictions inherent in any economic system: the underlying production relationships and market clearance for all goods and factors. In particular, those sectors of the market economy not subject to government intervention are assumed to be perfectly competitive, and the admissible set of government policy tools includes anything necessary to achieve a social welfare maximum. The government can redistribute any good or factor lump sum; it can change the price of any good or factor to consumers or firms; it can commander inputs and supply outputs at will, subject as always to given production relationships and market clearance; and it has perfect information about preferences, technologies, and markets. In short, the government has sufficient degrees of freedom to achieve Bator's bliss point, the social welfare maximum. It can design whatever policies are necessary to restore pareto optimality and bring society to its first-best utility−possibilities frontier. Then it can move society along the frontier to the bliss point by means of lump-sum redistributions that satisfy the interpersonal equity conditions.

Second-best analysis is a reaction to these heroic first-best assumptions. In an attempt to be more realistic, it posits at least one additional constraint on the policy environment. The constraint(s) can be on the underlying market environment, on the set of admissible government policy tools, or on the information available to the government. An immediate implication is that the search for the social welfare maximum covers a restricted set of allocations and distributions relative to first-best theory, illustrated by the shaded portion in Fig. 12.1. The defining difference from first-best analysis is that the restricted set cannot include point B, Bator's bliss point, because adding binding restrictions must reduce the maximum attainable level of social welfare. Whether or not any points on the first-best utility−possibilities frontier are feasible depends on the nature of the additional constraints, but such points might not be policy relevant anyway. As illustrated in Fig. 12.1, point A on U^2—U^1 is dominated by any point within the shaded portion and above the social welfare indifference curve W_1.

The entire thrust of second-best analysis is toward increased realism. For instance, second-best theory recognizes that governments cannot redistribute income lump sum. Taxes and transfers conditioned on income are almost always distorting. Similarly, all market economies contain some monopoly or monopsony elements that are unlikely to disappear in the foreseeable future. In addition, individuals and firms possess private information about themselves that others, including the government, do not know. Policy analysis should incorporate these real-world phenomena, which appear as additional constraints in a formal general equilibrium model.

As noted in Chapter 3, the most common government policy restrictions employed to date in the professional literature have been the inability to make lump-sum redistributions, the necessity of raising tax revenue in a distorting manner, the requirement that government agencies or entire governments operate within a legislated budget constraint, the not uncommon practice of governments either drafting some production inputs or offering some

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00012-1

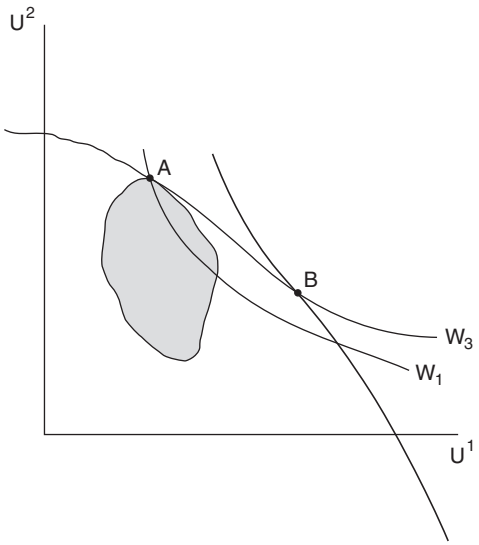

FIGURE 12.1

outputs of public projects free of charge (or at least at prices below opportunity costs), and the existence of private information. The most common market restriction assumes the existence of maintained monopoly power somewhere in the private sector, so that at least one private sector price does not equal marginal opportunity costs.

The potential set of policy, market, and information restrictions is limited only by the imagination, and individual constraints can always be combined, each additional constraint further restricting the set of feasible allocations and distributions. Thus, the possibilities for second-best analysis are virtually endless. There can never be *a* second-best normative theory of the public sector as there is with first-best theory.

Public sector economists thus face something of a dilemma in trying to inform public policy. They can recommend that policy makers try to approximate the definitive results from first-best theory, knowing that the underlying first-best model is patently unrealistic, or they can recommend that policy makers try to approximate the results from one or more second-best models, knowing that the results depend on the particular constraints that have been chosen in the name of realism and that the real world is many times more constrained than any one model can hope to capture.

Peter Hammond, in a brilliant review of the state of public sector economics published in 1990, came down firmly on the side of the second-best models. He urged public sector economists to dismiss out of hand the "delusions" of first-best theory, particularly its reliance on lump-sum redistributions (Hammond, 1990). He recommended instead that they push on with second-best theorizing if they wished to be taken seriously by policy makers, especially the most recent theory associated with private information. One wonders, however, whether Hammond believes that second-best theory will eventually achieve a conventional wisdom, an agreed-upon set of results that policy makers can rely on. His article gave no hint that second-best theory had yet come anywhere near this goal.

A BRIEF HISTORY OF SECOND-BEST THEORY

Second-best theorizing swept into public sector economics to stay in the 1970s and has been at the forefront of the discipline ever since. The impetus was provided by two seminal articles that appeared about 10 years apart: "The General Theory of Second-Best," by Archibald Lipsey and Kelvin Lancaster, and "Optimal Taxation and Public Production," by Peter Diamond and James Mirrlees.[1] The two articles approached second-best analysis from different perspectives and each became a template for distinct branches of the second-best literature that followed.

The Lipsey–Lancaster paper took the natural first step away from Samuelson's first-best general equilibrium model. Their model assumes that price exceeds marginal cost in at least one market, either because of private monopoly power or because of distorting government taxes, and that the government is unwilling or unable to remove the distortion. Otherwise, the government has as much freedom to act as it has in the first-best model, including the ability to redistribute income lump sum.

Lipsey and Lancaster were specifically interested in the following question: Given maintained distortions in some markets, are the first-best pareto-optimal rules for other markets still consistent with social welfare maximization? The answer turned out to be "no" in general, a result that became known as the "theorem of the second best." Subsequent research has expanded their analysis to consider the effects of maintained distortions on first-best public expenditure decision rules and on the welfare implications of changing the pattern of distorting taxes. One can ask, for example, whether substituting one set of distorting taxes for another while holding tax revenue constant increases social welfare, given the existence of still other distortions. These are issues of *policy reform* in a second-best environment, by now a huge body of literature in the Lipsey–Lancaster mold.

1. Lancaster and Lipsey (1956–1957); Diamond and Mirrlees (1971). The Diamond/Mirrlees paper was completed in 1968 and widely circulated as an MIT Working Paper. It was well known and widely cited by the time it was published in 1971. Two other early articles are worth mentioning, one on production and one on taxation: Boiteux (1971) and Stiglitz and Dasgupta (1971). These articles are frequently referenced in the second-best literature. For an excellent (but difficult) summary of second-best methodology, see Green (1975).

The Diamond–Mirrlees model was also only one step removed from Samuelson's first-best model, but they asked a different question from Lipsey–Lancaster. The only maintained restriction in their model is that the government must raise some tax revenue by means of distorting taxation; otherwise, the government is free to vary all price–cost margins, exactly as in first-best analysis. Because all tax rates are under the government's control, Diamond–Mirrlees were able to consider the optimal pattern of distorting taxation for raising a given amount of revenue. As such, their paper spawned another huge body of literature on *optimal policy* in a second-best environment. These two seminal papers, and the literature that has followed, have taken the normative analysis of tax and expenditure theory in many new directions.

Second-Best Tax Theory

The allocational theory of taxation, which analyzes the welfare losses caused by distorting taxes, dates from the very beginning of public sector economics. It has, by its very nature, always been part of the theory of the second best. The application of formal, second-best, general equilibrium models to tax problems over the past 40 years has mainly served to sharpen normative tax theory. The most notable extensions have been in the context of many-person economies. We now have a much better understanding of the trade-offs between equity and efficiency in a second-best environment. Tax theory has also become more tightly integrated with public expenditure theory.

Second-Best Expenditure Theory

The impact of second-best modeling on public expenditure theory has been nothing short of revolutionary. Until the 1960s, the received doctrine on public expenditures was the first-best theory of Part II. Since then, public expenditure theory has literally been rewritten by second-best theorizing. One might have anticipated this. Adding constraints to a general equilibrium model obviously changes its first-order conditions and the resulting policy decision rules. The changes have hardly been trivial, however. Second-best public expenditure decision rules often bear little relationship to their first-best counterparts, to the point that economists now seriously question the policy relevance of such cherished old "standards" as $\sum \text{MRS} = \text{MRT}$ for consumption externalities, or marginal cost pricing with subsidy for decreasing cost firms. Worse yet, it is now painfully obvious that the very latest, "state-of-the-art" second-best policy rules may not have much policy relevance either. As researchers invent new ways to constrain economic systems, they necessarily develop new and different, perhaps quite different, policy guidelines for the standard problems. As noted earlier, second-best theory is inherently a theory in flux, its policy implications always vulnerable to further variations in the models.

Second-best analysis has uncovered still other difficulties. To begin with, second-best public expenditure rules typically lack the comfortable intuitive appeal of the first-best rules. As we discovered throughout Part II, first-best policy rules always have close competitive market analogs. The correct price for decreasing-cost services is a pseudocompetitive price, and externality problems can be viewed as instances of market failure, meaning that a competitive market structure can always be described that will generate the correct pareto-optimal rules. These interpretations arise precisely because all first-best decision rules are simple combinations of marginal rates of substitution and transformation. Such is not the case for the second-best rules. The marginal rates of substitution and transformation are present, to be sure, but so are a number of terms embedded in the additional constraints that do not have natural competitive market interpretations or analogs. This is a discouraging outcome for believers in the competitive market system.

A second disappointment is that the first-order conditions of second-best general equilibrium models do not generally dichotomize into distinct sets of pareto-optimal and interpersonal equity conditions. Recall that first-best models dichotomize because the government is assumed to be able to lump-sum redistribute in order to satisfy the interpersonal equity conditions of social welfare maximization. In their quest for realism, second-best models usually deny the government that option, with the result that *all* second-best optimality conditions combine elements such as marginal rates of substitution (transformation), which appear only in first-best efficiency rules, *and* social welfare terms, which first-best theory isolates in the interpersonal equity conditions. Normative prescriptions such as "place a unit tax on each person's consumption of this particular good" tend to be replaced or modified by rules such as "tax those goods that are consumed relatively more by people with low social welfare weights." But, because we have no useful theory of interpersonal equity comparisons, these policy rules tend not to be terribly compelling. Economists can take some comfort in the knowledge that the second-best policy rules are useful to public officials so long as the officials are willing to provide the social welfare weights, but this is a far cry from having a complete normative theory of the public sector.

Economists sometimes avoid the social welfare terms altogether by resorting to the fiction of the one-consumer-equivalent economy. These models, however, can do little more than highlight the efficiency aspects of public sector problems. We will use one-consumer models for this purpose as well, but it should be understood that their policy implications are uncertain, unless it is assumed that

distributional equity has been achieved. This can occur only by chance, however, without lump-sum redistributions. (Alternatively, does anyone seriously believe that preferences are identical and homothetic, a sufficient condition for one-consumer equivalence?) Furthermore, given the likelihood of unequal social welfare weights, it is always possible to specify some pattern of weights such that the efficiency aspects of any given policy rule become relatively unimportant. Needless to say, the presence of the social welfare terms in the optimal decision rules for allocational problems such as externalities is extremely troublesome for normative public sector theory.

A final discouragement is that second-best restrictions tend to affect *all* markets, not just those in which public expenditures occur. First-best models have the property that government intervention in any one market does not change the form of the pareto-optimal rules for other goods and factors. They can be allocated in competitive, decentralized markets. The important implication is that instances of market failure can be corrected with policies targeted solely at the failure. This is no longer true in a second-best environment. The Lipsey–Lancaster theorem says that if price–cost margins are distorted in some markets, then first-best competitive efficiency rules are no longer optimal for other markets, in general. Roughly the same result applies in the Diamond–Mirrlees framework. If the government must raise revenue by means of distorting taxation, it is generally optimal to tax all goods and factors (except one). The thrust of second-best analysis, therefore, is toward pervasive rather than limited government intervention, a discouraging result indeed for decentralized capitalist economies and the government-as-agent principle of government intervention.

Private Information

The constraint that people possess private information about themselves that other people and/or the government do not know deserves separate mention in this brief history of second-best theory. Private information is also commonly referred to as asymmetric information. It has been one of the more important focal points of public sector analysis over the past 25 years, if not the most important. The intense interest in the implications of private or asymmetric information is understandable. It opens up a whole new range of possibilities for public sector economists to consider, possibilities that challenge much of the received doctrine in public sector theory.

Private information is different from the other second-best constraints because it is not simply a technological or practical assumption tacked on to an otherwise first-best model. It is in part an assumption about how people behave, that they are willing to use their private information for their own personal gain and to deceive if need be. As such, it leads normative public sector theory down a very slippery slope.

On the one hand, the idea that people are willing to deceive the government for their own ends tears at the very fabric of society. It belies the expectation of good citizenship and makes a mockery of the traditional notion that the government's proper economic function is as an agent of the people acting to correct market failures by pursuing the public interest in efficiency and equity. What is the normative appeal of maximizing an individualistic social welfare function when some people are willing to deceive and others are honest? Should the deceivers receive zero marginal social welfare weights? How much deception does it take before the society collapses? The objective function of public policy is not at all obvious when people are prone to act selfishly to exploit their private information.

On the other hand, some people certainly do use private information to their own advantage, and such behavior is entirely consistent with the economic view of individuals as self-interested utility maximizers. The willingness to exploit private information is not just a matter for positive economic analysis, however. It matters for normative analysis as well. All normative policy prescriptions must make assumptions about people's behavior and about how they will respond to the policies, and the prescriptions are only useful if the behavioral assumptions are reasonably accurate. Normative theory cannot simply ignore the issue of private information. The problem, though, is that the existence of private information can be extremely constraining for a government dedicated to the public interest in efficiency and equity, to social welfare maximization.[2]

The force of the private information constraint turns on the very meaning of an equilibrium in the social sciences, as a situation in which no one has any incentive to change his or her behavior. The particular requirement of an equilibrium in the presence of private information is that no one has any incentive to deceive or to represent one's private information as other than what it really is. The only feasible public policies are those that are consistent with this notion of equilibrium. To be feasible, therefore, a public policy must be such that everyone's best strategy is to tell the truth about themselves given the policy; deception cannot lead to personal gain. For example, high-income people cannot pretend to have low income in order to reduce their taxes.

2. The implications of private information for public sector analysis are masterfully set out in the overview article by Peter Hammond for the *Oxford Economic Papers*. He discusses the points raised here plus the implications of limited information for applied cost–benefit analysis. Hammond also references the most important journal articles in these areas. See Hammond (1990).

In the parlance of game theory, public policies must honor the revelation principle or, equivalently, be incentive compatible. In terms of formal modeling, private information necessitates adding one or more incentive compatibility constraints to a social welfare maximization problem to assure that the resulting policy prescription is feasible.

Incentive-compatibility constraints can indeed place severe restrictions on the set of feasible policies. To begin with, they rule out almost all lump-sum redistributions unless they can be targeted to readily observable characteristics that an individual cannot hide or change, such as age. The feasible redistributions are unlikely to have much distributional bite, however. In truth, the government really has no chance of satisfying the first-best interpersonal equity conditions in a world of private information. And, as we have seen, the entire body of first-best theory rests on shaky foundations when the scope of lump-sum redistributions is limited.

It turns out that economists have been unable to find very many public sector policies that are both efficient and equitable for which truth telling is the dominant strategy. The most obvious inventive-compatible distributional policy in the face of private information is to do nothing; simply accept the initial distribution of resources. This policy may be consistent with efficiency but it is likely to be seen as unjust.

Another variation of private information is the ability to engage in market exchanges in the underground, informal sector of the economy, out of sight of the government. The possibility of underground exchanges can severely limit the government's ability to do much of anything if escape to the informal sector is relatively easy. For example, the government may not be able to collect taxes or enforce sanctions against illegal activities.

Notice, too, how the presence of an underground economy changes the perception of markets. The traditional view of markets is that they are the best mechanisms yet devised for promoting efficient exchanges. The relatively few exceptions are the instances of market failure that require government intervention, such as nonexclusive goods or decreasing-cost services. Markets in the underground economy, even highly competitive markets there, are destructive to efficiency, however. They sharply constrain the feasible set of government policies that can be used to promote efficiency (and equity).

The existence of underground economies is hardly a trivial problem, even in the industrialized market economies. Friedrich Schneider, Andreas Buehn, and Claudio Montenegro attempted to measure the size of the underground economy in 162 countries from 1999 to 2007. They chose a narrow definition of an underground economy, one consisting of market transactions that would be legal if undertaken in the regular economy and included in national income and product but that go underground either to avoid paying taxes or to escape certain regulations such as minimum wage laws, safety standards, and various administrative procedures. They ignored other illegal activity and all barter activity. They estimated that the underground economy averaged 17.1% of the total economy over 8 years for the 25 Organisation for Economic Co-operation and Development (OECD) countries, with a maximum of 28.0% (Cyprus) and a minimum of 8.5% (Switzerland, with the United States next lowest at 8.6%). The ratios were generally much higher in the non-OECD countries.[3]

Still another variation of private information that causes problems for normative public sector theory is the limited information that consumers and the government have about their relevant opportunity sets. Regarding the consumers, traditional microeconomic analysis assumes that consumers maximize their utilities with full information about their opportunity sets, including perfect foresight about future events. In fact, consumers often have very limited information about their opportunity sets and little economic incentive to obtain much more information. Instead, they engage in some form of bounded rationality, often basing their decisions on simple rules of thumb consistent with the limited information available to them. Normative policy analysis typically assumes that consumers maximize under full information because it is the convenient assumption to make. If consumers instead use simple rules of thumb, questions arise as to what rules they follow, and how they change their behavior as their information sets change. There are no obvious answers to these questions, yet a normative theory has to know how consumers will respond to public policies. Regarding the government, public policies often result in large changes in the economy that affect many people and many prices. Policy makers are hard pressed to keep track of all the general equilibrium changes in the economy, to say the least. They, too, have only limited information, not enough to know for sure whether any given policy increases or decreases social welfare.[4]

In summary, the presence of private or asymmetric information offers any number of challenges to traditional normative public sector theory. The challenges are especially strong if private information takes a form that utility-maximizing individuals can use for their personal gain. For then the government has to be concerned with designing incentive-compatible policies that may severely limit its ability to pursue the public interest in efficiency and equity, which mainstream economists view as its primary function. At some point the willingness to deceive may so restrict the government's options that economic policy is hardly worth

3. Schneider et al. (2010). The definition of the underground or shadow economy is on p. 444, and the percentages listed are in Tables 2 and 3, pp. 454–457.

4. For an expanded discussion of the problems caused by limited information, see Hammond (1990).

doing. The social contract is broken and the goal of developing a normative public sector theory is no longer compelling.

We will demonstrate the implications of private information at various points in this part of the text. The underlying assumption throughout is that the government's pursuit of efficiency and equity remains a worthwhile endeavor.

PHILOSOPHICAL AND METHODOLOGICAL UNDERPINNINGS

Second-best theory shares the same philosophical and methodological foundations as first-best theory. The added constraints of second-best theory are the only important differences between them. For instance, consumer sovereignty remains the fundamental value judgment of second-best theory, and distributional considerations are most often represented by a Bergson–Samuelson individualistic social welfare function. Second-best analysis is also closely tied to the competitive market system. This is best illustrated by the observation that much of the second-best literature uses general equilibrium models expressed in terms of competitive market prices, not quantities. Analytical constructs such as indirect utility functions, expenditure functions, production functions expressed in terms of market supply (input-demand) functions, and generalized profit functions are commonplace in second-best analysis, and they all implicitly assume competitive market behavior.

There is an obvious reason why this has happened. The second-best literature has been centrally concerned with restrictions in the form of price–cost differentials, most often resulting from distorting taxation. Models already specified in terms of prices can incorporate these distortions more readily than models specified in terms of quantities. In turn, the easiest way to convert a general equilibrium quantity model into a price model is by assuming competitive price-taking behavior. Thus, in nearly all second-best models consumers are assumed to maximize utility subject to a fixed-price budget constraint. They have no market power. Producers are typically viewed as decentralized, perfectly competitive profit maximizers, often with simple production relationships exhibiting either constant costs or constant returns to scale. Even the government is assumed to transact at the competitive producer prices to the extent it buys and sells inputs and outputs. A second-best model might posit constraints in the form of noncompetitive behavior in a small subset of markets, but the underlying market economy is almost always competitive. Second-best results may not have competitive interpretations, but the majority of models used to date have been competitive through and through.

The newer literature on private information is somewhat of an exception because it applies the techniques of game theory, and the games being played may not occur in a market setting. Even so, the decentralized nature of the competitive market has a correspondence in the public sector allocation mechanisms that honor the revelation principle. A standard requirement of truth-revealing mechanisms is that individuals have no control over their opportunity sets. The public sector mechanisms must be decentralizable in this sense.

In summary, although second-best theory has severely challenged all first-best policy rules, it has taken only the smallest, most hesitant steps away from the highly stylized first-best policy environment. Second-best analysis is more realistic, but only slightly so.

PREVIEW OF PART III

With these reflections in mind, we will begin our second-best analysis with the allocational theory of taxation, thereby reversing the order of presentation in Part II. This happens to coincide with the historical development of second-best theory, but that is really beside the point. There are two good analytical reasons for considering tax theory first.

One is that second-best tax theory is inherently simpler than second-best expenditure theory, in this sense. Public expenditure theory requires the specification of a distinct problem (e.g., an externality) and one or more distinct constraints (e.g., distorting taxation), whereas tax theory requires only the specification of a constraint. Saying that all taxes must be distorting is at once an additional constraint on the system and the source of the problem being analyzed in tax theory. Consequently, problems in tax theory can be analyzed with much simpler general equilibrium models. This is an important advantage. Second-best models specified in terms of prices are quite different from the first-best quantity model of Part II, enough so that it pays to begin the analysis as simply as possible. Thus, the initial chapters on tax theory have two goals. Their main purpose is to demonstrate some important theorems in the allocational theory of taxation, but they also serve as an introduction to second-best methodology.

Second-best tax theory also logically precedes public expenditure theory, so long as distorting taxes are one of the policy constraints. Having studied the effects of distorting taxation in isolation, the implications for public expenditure issues such as externalities and decreasing-cost production are that much more apparent. Chapters 13–17 contain a detailed analysis of the theory of distorting taxes, often without any consideration of how governments actually spend tax revenues. Chapters 18–24 then rework selected public expenditure problems from Part II within a second-best framework—transfer payments to the poor, aggregate externalities, nonexclusive goods, decreasing costs—using the constraints most commonly employed in the literature. They also include an analysis of public

insurance. Chapter 25 concludes Part III with a discussion of behavioral economics, which studies behavior of individuals that is anomalous or irrational in the sense that it is clearly not utility maximizing. The chapter focuses on a few of the more important and widespread anomalies, highlighting the severe challenges they pose for mainstream public sector theory.

REFERENCES

Boiteux, M., September 1971. On the management of public monopolies subject to budgetary constraints. Journal of Economic Theory 3 (3), 219–240 (translation from French, Econometrica, January 1956).

Diamond, P.A., Mirrlees, J., March, June 1971. Optimal taxation and public production" (2 parts; part I: production efficiency and part II: tax rules). American Economic Review 61 (1), 8–27; 61 (3, Part 1 of 2), 261–278.

Green, H., November 1975. Two models of optimal pricing and taxation. Oxford Economic Papers 27 (3), 352–382.

Hammond, P., January 1990. Theoretical progress in public economics: a provocative assessment. Oxford Economic Papers 42 (1), 6–33.

Lancaster, L., Lipsey, R., 1956–1957. The general theory of second-best. Review of Economic Studies 24 (1), 11–32.

Schneider, F., Buehn, A., Montenegro, C., December 2010. New estimates for the shadow economies all over the World. International Economic Journal 24 (4), 443–461.

Stiglitz, J., Dasgupta, P., April 1971. Differential taxation, public goods, and economic efficiency. Review of Economic Studies 38 (2), 151–174.

Chapter 13

The Second-Best Theory of Taxation in One-Consumer Economies with Linear Production Technology

Chapter Outline

The second-best theory of taxation explores the effects of distorting taxes on social welfare. A distorting tax is one that prevents at least one of the first-best pareto-optimal conditions from holding, that is, it forces society inside its first-best utility—possibilities frontier. The first-best pareto-optimal conditions are equalities between marginal rates of substitutions and marginal rates of transformation. As such, they require that agents face the same prices in a market economy. Distorting taxes prevent the equalities from holding because they force at least two economic agents in the same market to face different prices in an otherwise perfectly competitive economy. Virtually all taxes actually employed by governments introduce some distortion into the economy, whether they be sales, excise, income, or wealth taxes (transfer payments are automatically included in the analysis because transfers are analytically equivalent to negative taxes).

Because tax distortion is defined relative to pareto optimality, much of the literature on second-best tax theory has treated it strictly as an allocational issue, concerned only with questions of economic efficiency. Consequently, the analysis often occurs within the context of one-consumer economies, a simplification that makes sense if one is willing to ignore distributional concerns. As Chapter 12 noted, however, second-best analysis has shown that allocational and distributional issues do not dichotomize in a second-best environment without lump-sum taxes and transfers, thereby raising questions about the policy relevance of considering the efficiency aspects of distorting taxes independently from their distributional consequences. Probably no one today would recommend a set of taxes simply on the basis of their efficiency properties. Nonetheless, it is analytically convenient to isolate the efficiency effects of taxes by using one-consumer economy models.

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00013-3

We can then consider the tax rules in many-person economies as combinations of efficiency and distributional elements, with the latter represented by the social welfare weights derived from an individualistic social welfare function. This is the approach we will take in developing the second-best theory.

The theory of distorting taxation addresses three main questions, one associated with welfare loss, another with optimality, and a third with tax reform:

1. *Welfare loss*: Relative to the first-best optimum, what is the loss in social welfare associated with any given set of distorting taxes (including a single tax)? Harold Hotelling provided the first rigorous analysis of this issue in his 1938 article, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates" (Hotelling, 1938). Arnold Harberger rekindled interest in this question in two separate articles appearing in 1964, "Taxation, Resource Allocation and Welfare" and "The Measurement of Waste."[1] By now the literature is voluminous, with these three articles standing as the seminal contributions.
2. *Optimality*: Relative to the first-best optimum, what pattern of distorting taxes minimizes the loss in social welfare for any given amount of tax revenue the government might wish to raise? This question has been explored under two separate assumptions: (1) that the government can tax all goods and factors and (2) that a subset of goods and factors must remain untaxed. The study of optimal taxation under the first assumption is commonly referred to as the optimal commodity tax problem, with seminal contributions by Frank Ramsey in "A Contribution to the Theory of Taxation" (1927), and Peter Diamond and James Mirrlees' "Optimal Taxation and Public Production" (1971).[2] Explorations of optimal taxation under the second assumption are a more recent phenomenon.[3] We will defer discussion of restricted optimal taxation until Chapter 15 when we consider an important subset of that literature, the theory of optimal income taxation. Mirrlees' "An Exploration in the Theory of Optimum Income Taxation" is the seminal article on optimal income taxation (Mirrlees, 1971).
3. *Tax reform*: Holding tax revenues (or the government budget constraint) constant, what is the change in social welfare from substituting one set of distorting taxes for another? Once again, the literature on this question is voluminous, with the seminal article by Corlett and Hague, "Complementarity and the Excess Burden of Taxation" (Corlett and Hague, 1953–1954).

Most of the formal analysis of these three questions employs general equilibrium models specified in terms of prices. Therefore, we will switch at this point from quantity models to price models in order to familiarize the reader with the most common second-best methodology.

GENERAL EQUILIBRIUM PRICE MODELS

General equilibrium price models can be rather complex or extremely simple depending upon the assumptions made regarding the nature of demand and the underlying production technology for the economy and whether the economy is static or dynamic. Choices on demand range from one-consumer-equivalent economies to many-person economies with interpersonal equity rankings determined by a Bergson–Samuelson social welfare function. The key choice with respect to production technology is whether production exhibits linear or general technology and, if the latter, whether or not the technology is constant returns to scale (CRS). The choice of production technology also has direct implications for the way in which market clearance is specified. Moving from static to dynamic analysis raises a whole new set of modeling issues, such as how different cohorts of people behave (e.g., the working young and the retired elderly), how people form expectations about the future, how asset markets clear as capital accumulates, and how technology changes over time.

In order to highlight the economic intuition of tax distortions, we will begin with the simplest possible general equilibrium model, a static one-consumer-equivalent economy with linear aggregate production technology. The one-consumer assumption removes all distributional considerations, so that welfare loss means efficiency loss and the theory of distorting taxation is purely an allocational theory. Positing a linear technology is enormously convenient because it exhibits constant marginal (opportunity) costs along the linear production–possibilities frontier. Since the economy is assumed to be perfectly competitive, this means that all relevant production parameters can be described by a vector of fixed producer prices, assumed equal to the constant marginal costs (or value of marginal products for factors). Furthermore, output supply (and input demand) curves are perfectly elastic at the fixed prices within the boundaries of the aggregate production frontier. Hence, market clearance is implicit because supplies (input demands) automatically expand or

1. Harberger (1964a,b). An excellent recent reference for the early literature is Green and Sheshinski (1979).
2. Ramsey (1927) and Diamond and Mirrlees (1971). Two excellent surveys of the optimal tax literature are Sandmo (1976) and Bradford and Rosen (1976). Finally, Diamond and McFadden (1974), contains an excellent analysis of some of the second-best tax issues analyzed in this chapter.
3. Dixit presents a lucid analysis of restricted taxation using the model to be developed in this chapter in Dixit (1975). Also see Dixit and Munk (1977).

contract to meet the consumer's demands (factor supplies). That is, output (and factor supply) is completely determined by the consumer's preferences at the given prices within the boundaries of the frontier. Market clearance is also irrelevant to the determination of prices, which are solely a function of the production technology.

The one main drawback to the linear technology assumption is that its simplicity tends to mask the role of production in determining the welfare costs of distorting taxes. Fortunately, however, a fair number of properties of distorting taxes do carry over virtually intact from linear to general technologies, especially if the latter exhibit CRS. In any event, we will relax the assumption of linear technology in Chapter 14.

THE MEASUREMENT OF LOSS FROM DISTORTING TAXES

The first question of distorting taxation concerns the measurement of welfare loss: Relative to the first-best optimum, what is the social welfare loss resulting from any given pattern of distorting taxes, within the context of a one-consumer, linear production economy? With one consumer, loss in social welfare is equivalent to the consumer's loss in utility. To be concrete, assume that the distorting taxes (transfers) take the form of unit taxes on both the consumer's purchases of goods and services and his supply of factors, levied on the consumers. In principle, then, the taxes include most forms of sales and excise taxes on the product side and income taxes on the factor side. In practice, only income taxes are typically paid by consumers; sales and excise taxes are paid by firms. As we shall discover, however, it makes no difference to loss measurement whether the government levies a tax on the demand or the supply side of any market. Therefore, the assumption that consumers pay a sales or excise tax is of no consequence. The only distorting taxes specifically ruled out at this point are the so-called partial taxes paid by certain firms (consumers) but not others, such as the corporation income tax.

The Geometry of Loss Measurement: Partial Equilibrium Analysis

The analysis of welfare loss from distorting unit taxes dates back to the beginning of public sector theory and has long appeared in economic texts at all levels, including introductory principles texts. Fig. 13.1 depicts the standard textbook analysis of the loss from a single tax under linear technology. S and D represent the zero-tax market supply and demand curves for a particular product. The no-tax equilibrium price is p, the constant supply price. A unit tax levied on the consumer shifts the demand curve down everywhere by the amount of the tax, to D′ in the figure.

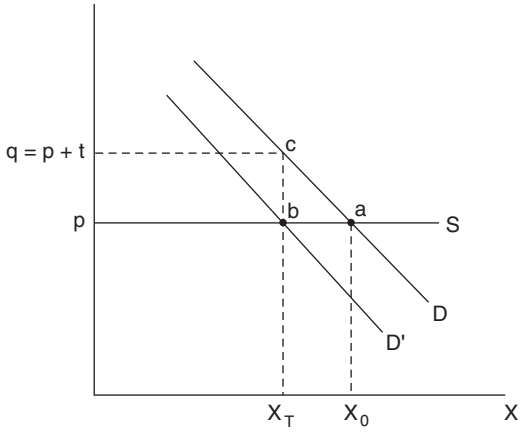

FIGURE 13.1

As a result of the tax, equilibrium output drops from X_0 to X_T, and the price to the consumer rises to $q = p + t$, where t is the unit tax. With constant producer prices, the price to the consumer rises by the full amount of the tax. The government collects revenue equal to $X_T \cdot t$.

The *deadweight* or *efficiency* loss of the tax is measured by the triangle abc, by the following argument. Because the tax causes the consumer price to rise from p to q, the consumer loses Marshallian consumer surplus in amount equal to the trapezoidal area qpac. Some of this loss is captured by the government as revenue $X_T \cdot t$, the rectangle qpbc, which presumably is used to finance socially beneficial expenditures. But the loss of triangle abc is captured by no one. It is a pure or deadweight welfare loss, generated by the distorting tax that forces consumers and producers to face different prices for the product. The consumer price q is called the *gross-of-tax* price, and the producer price p is the *net-of-tax* price. The loss triangle is an indication that the pareto-optimal condition MRS = MRT no longer holds in this market. Consumers equate their MRS to q, and producers equate their MRT to p (relative to the numeraire good). Without the tax, both the MRS and MRT are equated to p.

The traditional analysis is intuitively instructive, but it is not a valid general equilibrium presentation of the loss question. We saw in Chapter 9 that Marshallian consumer's surplus is not a meaningful compensation measure of loss, in general.[4] Moreover, it can be seriously misleading. For instance, one "theorem" commonly derived from the supply and demand framework is that the government should tax products (factors) whose demand (supply) is perfectly inelastic to avoid deadweight loss. If either the demand or supply curve in Fig. 13.1 were vertical, the output would remain constant, and there would be no deadweight loss triangle resulting from the unit tax. Unfortunately, this

4. Chapter 9 has a detailed discussion of this point.

proposition is not accurate. Unit taxes can generate welfare loss, properly measured, even if demand or supply is perfectly inelastic.

Another limitation of the standard textbook discussion is that it is a partial equilibrium analysis. As such, it cannot capture the effects on loss of further price changes in other markets.

The Geometry of Loss Measurement: General Equilibrium Analysis[5]

The first task, then, is to develop a proper and unambiguous measure of the welfare loss resulting from distorting taxes in a full general equilibrium context. To capture the intuition behind the measure, we will continue with a graphical analysis but switch from the partial equilibrium supply–demand framework to a valid general equilibrium representation using the consumer's indifference curves and the economy's production–possibilities frontier.

Suppose the consumer buys a single good, X_1, and supplies a single factor, X_2 (e.g., labor, measured negatively), with preferences $U(X_1, X_2)$, represented by the indifference curves I_1, I_2, and I_3 in Fig. 13.2. In addition, assume that producers can transform X_2 into X_1 according to the linear technology $X_1 = a \cdot X_2$, where a is the marginal product of X_2 (labor). The production–possibilities frontier is depicted as line segment Ob in Fig. 13.2. All feasible (X_1, X_2) combinations lie on or to the southwest of Ob. Given the consumer's preferences and the economy's production possibilities, point A is the first-best welfare optimum for the economy. Point A can be achieved by a competitive equilibrium, with relative prices P_{X_2}/P_{X_1} equal to the slope of the production frontier. That is, $P_{X_1} = P_{X_2}/a = MC_{X_1}$, the standard competitive result. To see that this is a general market equilibrium, note that with $P_{X_2}/P_{X_1} = a$, the production–possibilities frontier Ob is also a budget line for the consumer, with zero lump-sum income (payment):

$$\frac{P_{X_2}}{P_{X_1}} = a = \frac{X_1}{X_2} \tag{13.1}$$

$$P_{X_1} \cdot X_1 = P_{X_2} \cdot X_2 \tag{13.2}$$

Thus, the consumer can purchase the optimal bundle (X_2^A, X_1^A). Furthermore, Ob represents the profit function for the firm with competitive pricing and indicates that the firm just breaks even. There are no pure economic profits (losses) to distribute to the consumer. Thus, Eqn (13.2) holds for both the consumer and the producer, and point A is the pretax competitive general equilibrium for the economy.

5. The analysis in this section draws heavily on unpublished class notes provided to us by Professors Peter Diamond and Paul Samuelson of MIT. See also Diamond and McFadden (1974).

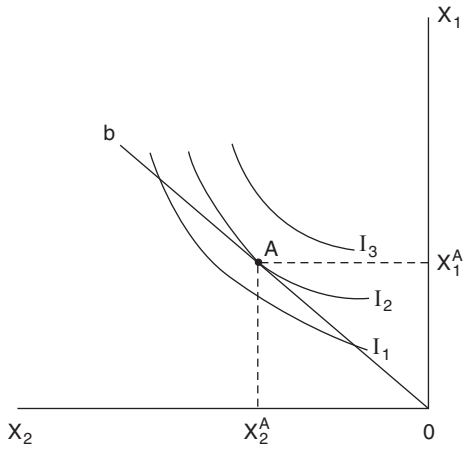

FIGURE 13.2

The first point to stress is that *any* tax on the consumer generates a loss in utility. Suppose the government places a lump-sum tax on the consumer in an amount equivalent to T_1 and forces him to a new equilibrium, point B in Fig. 13.3. Clearly the utility at B is less than the utility at A; the consumer has suffered a loss. Yet because lump-sum taxes are nondistorting, they cannot possibly generate a deadweight loss. Hence, the loss $U(\text{A}) - U(\text{B})$ must be considered an unavoidable consequence of any tax and should not be included in the measure of loss arising from tax distortion.

To see that a distorting tax generates loss in addition to this unavoidable loss, place a unit tax, t_1, on the consumption of X_1 such that it raises the same amount of revenue as the lump-sum tax. This tax changes the relative prices faced by the consumer from P_{X_2}/P_{X_1} to $P_{x_2}/(P_{x_1} + t_1)$, while leaving the relative producer prices at P_{X_2}/P_{X_1}. Since the consumer and producers now face different relative prices, $\text{MRS}_{X_1,X_2} \neq \text{MRT}_{X_1,X_2}(\equiv \text{MP}_{X_2}^{X_1})$, pareto optimality cannot

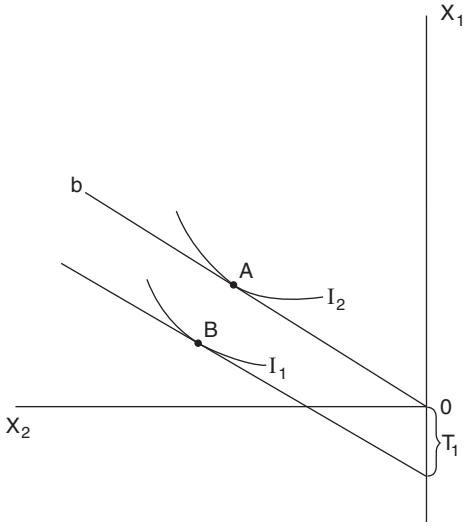

FIGURE 13.3

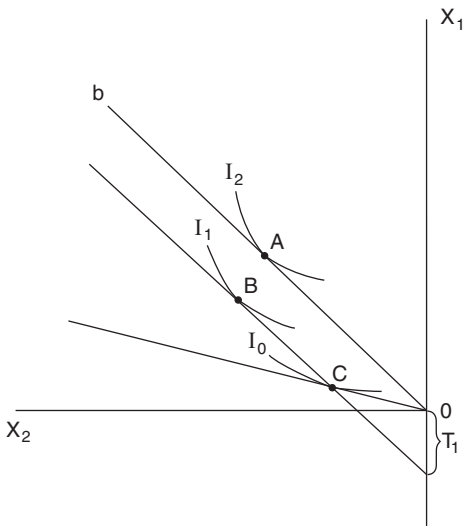

FIGURE 13.4

obtain, and the tax is distorting, by definition. The consumer chooses a new equilibrium, point C in Fig. 13.4.

Notice that the lump-sum and unit taxes raise the same amount of revenue, T_1 units of X_1, but that the equilibrium points B and C differ.[6] In general, C will be to the southeast of B, as drawn. The reason is that the unit tax introduces a *substitution effect* because of the relative price change that is absent from the lump-sum tax. The *income effects* of the two taxes are identical by construction, each measured by the lost tax revenue, but the added substitution effect causes the consumer to purchase less X_1 and less X_2 than with the lump-sum tax (more negative X_2, the "good"—for example, leisure). Also, C provides less utility than B, as drawn. This follows from revealed preference. When the consumer purchased B with the lump-sum tax, he was able to purchase C. Conversely, when he purchased C with the unit tax, he was unable to purchase B. Hence, B is revealed preferred to C. Only if the indifference curves are right angled, so that there is no substitution effect with the unit tax, do the two taxes generate the same after-tax equilibrium and thus the same loss in utility.

The additional loss in utility from B to C, then, can be considered the *avoidable loss* of the distorting unit tax, the loss corresponding to the deadweight loss triangle in the supply and demand presentation of Fig. 13.1. This is the loss society is interested in minimizing. Furthermore, the graphical analysis suggests that the amount of the avoidable loss for any distorting tax depends upon two factors: (1) the level of the tax rate and (2) the magnitude of the substitution effects between goods and factors.

6. It is always possible to construct equal-revenue taxes by positing any given unit tax, finding the new equilibrium, and constructing a line through this equilibrium parallel to the no-tax budget line to represent the equivalent lump-sum tax.

Of course, we would not want to measure the avoidable loss as the difference in utility levels $U(\mathrm{B}) - U(\mathrm{C})$, since this measure is not invariant to monotonic transformations of the utility index. What is required is an unambiguous income measure of the avoidable utility loss. As discussed in Chapter 9 when considering the "hard case" for decreasing cost services, such income measures involve the notion of compensation or willingness to pay. There is an infinite number of acceptable income compensation measures because they are all based on parallel distances between indifference curves, which can be computed at an infinity of points. One particularly intuitive income measure of tax loss is obtained from the following conceptual experiment:

1. Place a unit tax on the consumer's purchases of one of the goods or factors (say, X_1).
2. Simultaneously transfer to the consumer enough income, lump sum, to keep him on the original zero-tax indifference curve.
3. Include in this income the tax revenue collected from the unit tax.
4. Ask if the tax revenue alone is sufficient compensation. If not, then measure the loss as the difference between the lump-sum income necessary to compensate the consumer and the tax revenue collected and then returned lump sum.

Because utility is being held constant at the original no-tax equilibrium and the income computed at the with-tax price, this measure utilizes Hicks' Compensating Variation resulting from the (relative) price change.

Consider first the lump-sum tax by this measure of loss. No matter what the size of the tax, if the consumer receives the revenue back lump sum simultaneously as it is collected from him, he remains at the original no-tax equilibrium. No further lump-sum income is necessary as compensation, and the loss measure is zero, as it must be. With the unit tax, however, the tax revenue is not sufficient compensation and the loss measure is positive, providing that the substitution effect is nonzero. This case is illustrated in Fig. 13.5.

Given the lump-sum compensation, the consumer remains on indifference curve I_2 as the budget line rotates in response to the tax. Suppose the consumer winds up at point D. D is a compensated market equilibrium in which the consumer faces the with-tax price line HE and simultaneously receives income lump-sum equal to 0E (in terms of X_1). The tax revenue collected (and returned) at the compensated equilibrium D equals EF units of X_1, the difference between the no-tax and with-tax price lines at D projected back to the X_1 axis. Hence, the distance OF measures the loss, the income (in units of X_1) required in excess of the tax revenue to compensate the consumer for the tax. OF is positive as long as the tax generates a substitution effect.

Note, finally, that society cannot produce the compensated equilibrium D because it lies outside the production—possibilities frontier Ob. This is another useful

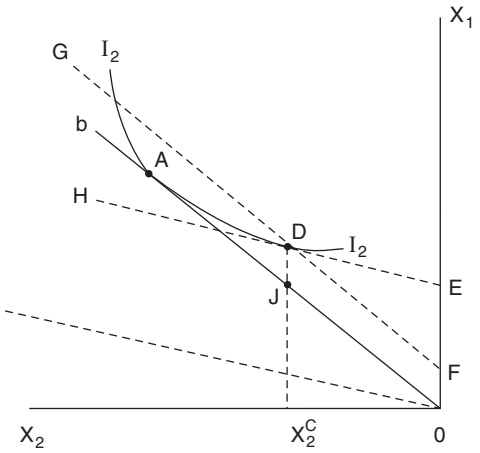

FIGURE 13.5

way of conceptualizing the notion of deadweight or avoidable loss, that society cannot satisfy the entire set of compensated goods demands and factor supplies from its own resources given distorting taxes. Suppose the consumer supplies the amount of labor at the compensated equilibrium, X_2^C. Producers can only supply J units of X_1 on 0b given X_2^C. Hence, the vertical distance (D−J) provides an alternative measure of loss, equal to OF, the amount of X_1 that the government would have to obtain from an outside source to compensate the consumer fully for the unit tax.[7]

It should be understood that these compensation experiments are merely conceptual exercises, useful for deriving certain analytical properties of distorting taxes. They are not indicative of actual government policies. There is no reason to suppose that a compensated equilibrium would ever actually be observed, and normative theory surely does not require that governments simply return whatever taxes they collect lump sum.

The Analytics of General Equilibrium Loss Measurement

Extending the concept of efficiency loss from distorting taxation to N goods and factors and any given pattern of existing taxes can be accomplished quite easily by means of the expenditure function from the theory of consumer behavior:

$$M(\vec{\mathbf{q}}; \overline{U}) = \sum_{i=1}^{N} q_i X_i^C(\vec{\mathbf{q}}; \overline{U})$$

where

$\vec{\mathbf{q}}$ = the vector of consumer prices, with element q_i and $X_i^C(\vec{\mathbf{q}}); \overline{U}$ = the compensated demand (supply) for good (factor) i.

$M(\vec{\mathbf{q}}); \overline{U}$ gives the lump-sum income for any vector of consumer prices necessary to keep the consumer at utility level $\overline{U}$. But, if $\overline{U}$ is set equal to the original zero-tax utility level, that is, $\overline{U} = U^0$, then $M(\vec{\mathbf{q}}; \overline{U}^0)$ is precisely the income measure required for the conceptual experiment described above.[8] Furthermore, there can be no pure profits or losses in the economy with linear technology. Hence, it is reasonable to assume that

$$M\left(\vec{\mathbf{p}}; \overline{U}^0\right) \equiv 0$$

where:

$\vec{\mathbf{p}}$ = the vector of producer prices, assumed fixed and equal to marginal costs.

With the zero-profit assumption, the loss for any given tax vector is the value of the expenditure function at the gross-of-tax consumer price vector less the tax revenues collected and returned (conceptually) lump sum, or

$$L(\vec{\mathbf{t}}) = M\left(\vec{\mathbf{q}}; \overline{U}^0\right) - \sum_{i=1}^{N} t_i \cdot X_i^C\left(\vec{\mathbf{q}}; \overline{U}^0\right) \qquad (13.3)$$

where

$\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$, and $\vec{\mathbf{t}}$ is the vector of unit taxes, with element t_i.

The tax revenue is the only source of lump-sum income available to the consumer.

Notice that the tax revenue is the revenue that would be collected at the fully compensated equilibrium, corresponding to point D in Fig. 13.5. To be consistent, the conceptual experiment must assume that compensation is actually paid, in which case the tax revenues collected from the vector of rates $\vec{\mathbf{t}}$ is $\vec{\mathbf{t}} \cdot X(\vec{\mathbf{q}}; \overline{U}^0)$. Actual tax collections, equal to $\sum_{i=1}^{N} t_i X_i(\vec{\mathbf{q}}; 0)$ where $X(\vec{\mathbf{q}}; 0)$ represents the consumer's ordinary or Marshallian demand (supply) curves, are irrelevant to this conceptual loss experiment.

7. Alternatively, the dotted line GF represents a production−possibilities frontier in which producers receive a lump-sum transfer of units of X_1 from an outside source. Given this transfer, and with the ratio of *producer* prices $P_{X_2}/P_{X_1} = a$, competitive production can achieve the compensated equilibrium D.

8. It should be noted that the choice of $\overline{U}$ is arbitrary since any constant utility level generates the same analytical expressions for total and marginal loss. Setting $\overline{U} = U^0$ is a natural choice when measuring the loss from distorting taxation, since loss is then defined explicitly with reference to the zero-tax, nondistorted economy. As noted in the text, setting $\overline{U} = U^0$ coincides with the conceptual loss experiment described above. Another intuitive choice would be to set $\overline{U}$ equal to the utility level obtained with lump-sum taxation. This would correspond to our introductory discussion of loss as represented in Fig. 15.4, in which loss is defined in terms of $U(C)$ versus $U(B)$, two equal tax revenue equilibria.

Relating Eqn (13.3) to Fig. 13.5, $M(\vec{\mathbf{q}}; U^0)$ corresponds to the distance 0E, $\sum_{i=1}^{N} t_i X_i^{\mathrm{C}}(\vec{\mathbf{q}}; \overline{U}^0)$ corresponds to the distance EF, and $L(\vec{\mathbf{t}})$ corresponds to the distance OF.

Note before proceeding further that the expenditure function $M(\vec{\mathbf{q}}; \overline{U}^0) \equiv M(\vec{\mathbf{p}} + \vec{\mathbf{t}}; \overline{U}^0)$ is, *by itself*, a valid general equilibrium model of a one-consumer economy with linear technology, when coupled with the standard assumption of perfectly competitive markets. On the demand side, the expenditure function incorporates all relevant aspects of the consumer's behavior.[9] On the supply side, the price vector $\vec{\mathbf{p}}$ specifies all relevant production parameters, since relative producer prices equal marginal rates of transformation with perfect competition. Market clearance is implicit. It is understood that supplies (input demands) respond with perfect elasticity to the consumer's demands (factor supplies) at the specified price vector $\vec{\mathbf{p}}$ and that the consumer automatically supplies all factors used in production and receives all the goods produced. Also, the resource limitations defining the outward limits to these supply responses depend entirely on the consumer's willingness to supply factors, which is already incorporated in $M(\vec{\mathbf{q}}; \overline{U}^0)$. Finally, once $\vec{\mathbf{t}}$ is set by the government, $\vec{\mathbf{q}}$ is determined by the relationships $\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$. Separate market clearance equations are not needed to determine equilibrium price vectors. Thus, given that $M(\vec{\mathbf{q}}; \overline{U})$ is a valid general equilibrium specification of a one-consumer economy with linear technology, it follows immediately that Eqn (13.3), along with the relations $\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$, is a valid general equilibrium specification of the conceptual loss experiment described in the preceding section.

Marginal Loss

As a first step in determining the policy implications of distorting taxation, consider the marginal loss from a small change in one of the unit taxes, t_k, all other taxes held constant:

$$\frac{\partial L(\vec{\mathbf{t}})}{\partial t_k} = \frac{\partial M\left(\vec{\mathbf{q}}; \overline{U}^0\right)}{\partial t_k} - \frac{\partial\left[\sum_{i=1}^{N} t_i X_i^{\mathrm{C}}\left(\vec{\mathbf{q}}; \overline{U}^0\right)\right]}{\partial t_k} \qquad k = 1, \ldots, N \tag{13.4}$$

$$\frac{\partial L(\vec{\mathbf{t}})}{\partial t_k} = M_k - M_k - \sum_{i=1}^{N} t_i M_{iK} \quad k = 1, \ldots, N \tag{13.5}$$

$$\frac{\partial L(\vec{\mathbf{t}})}{\partial t_k} = -\sum_{i=1}^{N} t_i M_{iK} \quad k = 1, \ldots, N \tag{13.6}$$

9. The compensated demands (factor supplies) come from the dual of the consumer's utility maximization problem: minimize expenditures, $\sum_{i=1}^{N} q_i X_i$, subject to utility being held constant (at U^0).

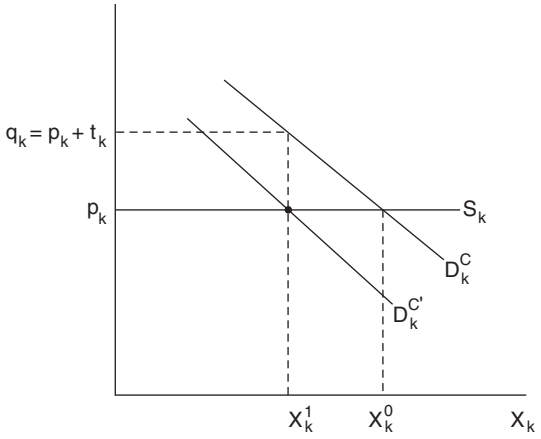

FIGURE 13.6

where:

$M_k = \frac{\partial M}{\partial q_k}$

$M_{ik} = \frac{\partial X_i^{\mathrm{C}}}{\partial q_k}$, the substitution terms in the Slutsky equation.

The derivatives take on these values because of the assumption of linear technology, which fixes the vector of producer $\vec{\mathbf{p}}$. In the kth market, represented by Fig. 13.6, the demand curve shifts down by the amount of the tax, and the consumer price q_k increases by the full amount of the tax in the new equilibrium. Thus, $\mathrm{d}q_k = \mathrm{d}t_k$, with p_k constant. The change in t_k may well affect demand (factor supply) in some other markets, say, the market for good j, as represented in Fig. 13.7. Output increases from X_j^0 to X_j^1, but there is no change in the equilibrium price q_j, since neither p_j nor t_j can change as t_k is varied. p_j is constant because technology is linear, and t_j is a control variable for the government, assumed constant. Hence, the derivative $\partial M/\partial t_k = \sum_{i=1}^{N} (\partial M/\partial q_i)(\partial q_i/\partial t_k)$ contains the single term $(\partial M/\partial q_k)(\partial q_k/\partial t_k) = (\partial M/\partial q_k) = M_k$; similarly, $\partial X_i^{\mathrm{C}}(\vec{\mathbf{q}}; \overline{U}^0)/\partial t_k = M_{ik}$.

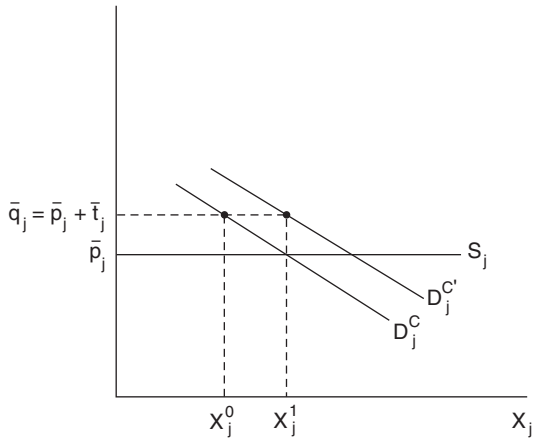

FIGURE 13.7

Notice that Eqn (13.6) for marginal loss confirms the results suggested by the one-good, one-factor graphical analysis: the additional loss from a marginal change in a distorting tax depends only upon the level of taxes already in existence and the Slutsky substitution effects between all pairs of goods and factors, M_{ik}.

Total Loss for Any Given Pattern of Taxes

Equation (13.3) gives one valid general equilibrium measure of the total loss from any given vector of distorting taxes, $\vec{\mathbf{t}}$. An alternative expression for total loss can be derived by integrating the N expressions for marginal loss and summing over all markets:

$$L(\vec{\mathbf{t}}) = \sum_{i=1}^{N} \int_0^{t_i} \frac{\partial L}{\partial t_i} \mathrm{d}t_i = \sum_{i=1}^{N} \int_0^{t_i} \frac{\partial L(q_1, \ldots, q_{i-1};\ s;\ p_{1+1}, \ldots, P_N)}{\partial s} \mathrm{d}s \tag{13.7}$$

Substituting Eqn (13.6) into (13.7) yields:

$$L(\vec{\mathbf{t}}) = \sum_{i=1}^{N} \sum_{j \le i} \int_0^{t_i} -t_j M_{ij} \mathrm{d}t_i \tag{13.8}$$

The inside summation before the integral sign of Eqn (13.8) indicates that the taxes are being introduced market by market. Thus, $t_j = 0, j > i$. Moreover, since $M_{ij} = M_{ji}$ from the symmetry of the Slutsky substitution terms, the order of integration is irrelevant. That is, the order in which the government actually levies the given vector of taxes does not affect the value of the total welfare loss resulting from the entire set of taxes. Equation (13.8) is an exact measure of total welfare loss for a one-consumer economy with linear technology. It can be related to the standard geometric representation of deadweight loss triangles, properly measured using compensated demand curves, if the compensated demand derivatives, $M_{ik} = \partial X_i^{\mathrm{C}}(\vec{\mathbf{q}}; \overline{U}^0)/\partial q_k$, are assumed constant—that is, if the compensated demand curves are assumed to be linear over the relevant range of prices. With this assumption, M_{ij} can be taken outside the integrals so that Eqn (13.8) becomes:

$$L(\vec{\mathbf{t}}) = \sum_{i=1}^{N} \left(-\sum_{j=1}^{i-1} t_j M_{ji} \int_0^{t_i} \mathrm{d}t_i - M_{ii} \int_0^{t_i} t_i \mathrm{d}t_i \right) \tag{13.9}$$

Performing all N integrations yields:

$$L(\vec{\mathbf{t}}) = -\sum_{i=1}^{N} \left(\sum_{j=1}^{i-1} t_j t_i M_{ji} + \frac{1}{2} M_{ii} t_i^2 \right) \tag{13.10}$$

Rearranging terms:

$$L(\vec{\mathbf{t}}) = -\frac{1}{2} \sum_i \sum_j t_i t_j M_{ij} \tag{13.11}$$

Arnold Harberger first derived an expression of this form in his 1964 articles "Taxation, Resource Allocation and Welfare" and "The Measurement of Waste."[10]

To relate this expression to deadweight loss triangles, rewrite Eqn (13.11) as:

$$L(\vec{\mathbf{t}}) = -\frac{1}{2} \sum_{i=1}^{N} t_i \sum_{j=1}^{N} t_j M_{ij} = -\frac{1}{2} \sum_{i=1}^{N} t_i \sum_{j=1}^{N} t_j \frac{\Delta X_i^{\mathrm{C}}}{\Delta q_j} = -\frac{1}{2} \sum_{i=1}^{N} t_i \Delta X_i^{\mathrm{C}} \tag{13.12}$$

under the assumptions of constant M_{ij} and $\mathrm{d}q_j = \mathrm{d}t_j$ for a linear technology economy. Equation (13.12) appears to suggest that the total loss from a given vector of taxes can be approximated as the sum, over all markets, of deadweight loss triangles in each market, as taxes are added one by one. This is misleading, however, since the quantity base of these triangles, the ΔX^{C} in Eqn (13.12), represents the total general equilibrium change in each X_i in response to the *entire set* of tax distortions t_j, for $j \le i$. Thus, it is not correct to sum deadweight loss triangles as they are traditionally presented in partial equilibrium analysis, even with the proper compensated demand curves.

Consider a two-tax example in which the imposition of t_1 precedes the imposition of t_2.[11] As t_1 is imposed, the loss at that point is correctly approximated by the shaded triangle in Fig. 13.8. D_1 and D_1' are the pre- and posttax compensated demand curves for X_1 (at the zero-tax utility level) under the assumption that $q_2 = p_2$, the producer price of good 2. D_2 may shift in response to t_1, but with no resulting addition to (subtraction from) loss since its price equals marginal cost (the same is true of all other goods). Hence, loss is properly measured as:

$$\frac{1}{2} t_1^2 \frac{\partial X_1^{\mathrm{C}}}{\partial q_1} = \frac{1}{2} \cdot t_1 \cdot \frac{\partial X_1^{\mathrm{C}}}{\partial q_1} \cdot \Delta q_1 = \frac{1}{2} \cdot t_1 \cdot \Delta X_1^{\mathrm{C}}$$

the shaded triangle. It immediately follows that the traditional representation of loss as a triangle on a demand and supply diagram is accurate for a single tax, providing

10. See Harberger (1964a,b). Harberger refers to the Slutsky substitution terms specifically in each article, but only as special cases. Generally, his $\partial X_i/\partial q_j$ refer to the general equilibrium response of the X_i along the production–possibilities frontier, not the movement along the consumer's zero-tax indifference curve. See Chapter 26 for additional discussion. Harberger clarifies the meaning of his "demand" derivatives in Harberger (1971) and Harberger (1974).

11. Harberger presents a similar geometric analysis of adding losses across markets in Harberger (1964a).

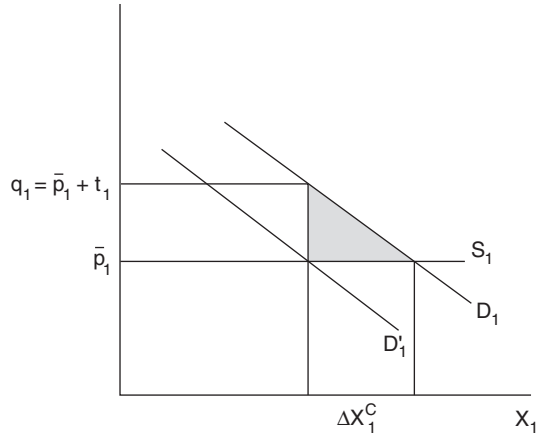

FIGURE 13.8

the compensated demand curves are employed and they are linear.

When t_2 is imposed, it creates an additional loss in the market for good 2, which can be represented by the shaded triangle in Fig. 13.9. The compensated pre- and posttax demand curves D_2 and D'_2 assume that $q_1 = p_1 + t_1$, the gross-of-tax price for good 1. The triangle equals $\frac{1}{2} \cdot t_2 \cdot M_{22} \cdot t_2 = \frac{1}{2} t_2 \cdot \Delta X_2^C$. However, if we simply add this loss triangle to the loss triangle in Fig. 13.8 for good 1 and stop, we will have ignored a third term in the expression for loss in Eqn (13.11) or (13.10), equal to $(-)t_1 \cdot t_2 \cdot M_{12}$, or $t_1 \cdot \partial X_1^C / \partial q_2 \cdot \Delta q_2$. Given that t_1 exists, the response of X_1 to a change in the price of X_2 entails a further source of loss since price no longer equals marginal cost for good 1. This additional loss equals the change in tax revenue collected from good 1 as its demand shifts. Recall that loss is the income required to compensate for the taxes less any tax revenue available for compensation. If the tax revenue collected in other markets changes as a new tax is imposed, there is more or less revenue available for compensation. Thus, if demand shifts to D''_1 as depicted in Fig. 13.10 (the two goods are Slutsky complements), then the shaded rectangular area (abcd) must be added to the standard deadweight loss triangle (cde) to complete the total loss associated with the market for good 1 resulting from the entire set of taxes t_1 and t_2. Note that D''_1 assumes $q_2 = p_2 + t_2$, whereas $D'_1(D_1)$ assumes $q_2 = p_2$. Thus, total loss from both markets is the trapezoidal area abed in Fig. 13.10 plus the triangle (fgh) in Fig. 13.9.

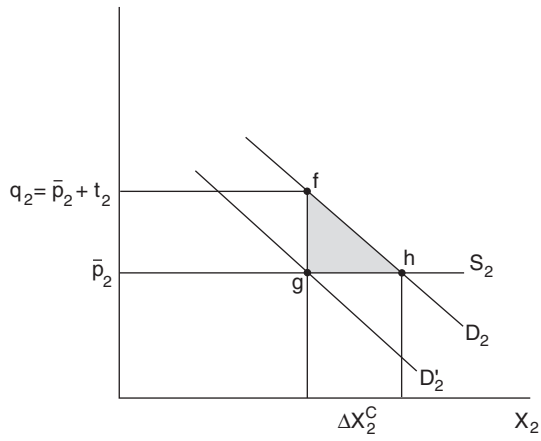

FIGURE 13.9

X_1 and X_2 could be Slutsky complements if there are more than two goods. They must be Slutsky substitutes ($M_{ij} > 0$) if they are the only two goods, however, from the homogeneity of the M_i. In this case, D'' shifts to the right of D' and the resulting rectangle represents a reduction in loss. The government collects more tax revenue from X_1 as t_2 is imposed, and the additional tax revenue is available to compensate the consumer for the increase in q_2. The additional tax revenue reduces the total loss.

The geometric analysis generalizes directly to N goods (and factors) in which rectangles of the form $t_i \cdot \Delta X_i$ in markets for which taxes already exist are added (subtracted) to the standard deadweight loss triangles $^1/_2\, t_k \cdot \Delta X_k$ as taxes t_k are added one by one. The triangles correspond to the terms $-\sum_i \frac{1}{2} M_{ii} t_i^2$ in Eqn (13.10); the rectangles, to the terms $-\sum_{i=1}^{N} \sum_{j=1}^{i-1} t_j t_i M_{ji}$.[12]

12. The loss measure (Eqn 13.11) can be directly related to our earlier discussion of the gain or loss to the consumer for any given change in consumer prices. In Chapter 9 we showed that the gain or loss for any price change can be represented as a summation of areas behind the consumer's compensated demand (and supply) curves between the old and new prices in each market. The result followed from the fact that

$$M(q^1; \overline{U}) - M(q^0; \overline{U}) = \sum_{i=1}^{N} \int_{q_i^0}^{q_i^1} \frac{\partial M(q_1^1, \ldots, q_{i-1}^1; s; q_{i+1}^0, \ldots, q_N^0; \overline{U})}{\partial s} \mathrm{d}s \tag{13.1n}$$

$$= \sum_{i=1}^{N} \int_{q_i^0}^{q_i^1} X_i^C(q_1^1, \ldots, q_{i-1}^1; s; q_{i+1}^0, \ldots, q_N^0; \overline{U}) \mathrm{d}s \tag{13.2n}$$

where $X_i^C(\overrightarrow{\mathbf{q}}^1; s; \overrightarrow{\mathbf{q}}^0; \overline{U})$ is the demand for X^i compensated at utility level $\overline{U}$ and evaluated at $q_j = q_j^1$, for $j < i$, and $q_j = q_j^1$, for j > i. With $\overline{U} = U^0$, these areas measure Hicks compensating variation. In the tax problem, $t_i = q_i^1 - q_i^0$ and $M(\overrightarrow{\mathbf{q}}; \overline{U}) = 0$, so the loss measure from distorting taxes corresponds directly to this earlier loss measure. The original measure gives the entire area behind each demand curve—for example, area $P_1 e d q_1$ in Fig. 13.10. As such, it captures only the change in the value of the expenditure function in response to the tax, that is, it ignores the disposition of the tax revenue. The tax loss measure recognizes that the revenue $P_1 b a_{q1}$ can be put to some socially useful purpose. Conceptually, it is simply returned lump sum to the consumer. Hence, the net or deadweight loss caused by the distortion is just the trapezoidal area abed.

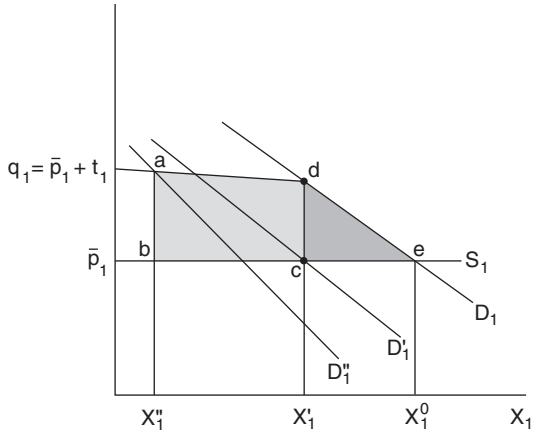

FIGURE 13.10

Policy Implications of the Loss Measures

The expressions (13.6) and (13.11) for marginal and total loss reveal a number of important policy implications on distorting taxation, despite the simplicity of the one-consumer, linear technology model. They all follow from the property that both marginal and total loss depend only upon the level of existing tax rates and the Slutsky substitution effects. Here are eight such implications as a representative sampling.[13]

Zero-Tax Economy versus Existing-Tax Economy

An immediate implication from the expression for marginal loss, Eqn (13.6), is that if there are no tax distortions in the economy, then the imposition of a marginal tax on one of the goods or factors does not generate a deadweight loss even to a second order of approximation. The level of all tax rates is either exactly or approximately equal to zero near the initial no-tax equilibrium so that marginal loss is also (approximately) zero. In other words, the first marginal distortion is free. The intuition behind this result is that all resource transfers in response to the new marginal tax occur at values (approximately) equal to their marginal costs. If so, then returning the tax revenue is sufficient compensation for the distortion.

Of course, the zero-tax, zero-loss result is just a theoretical *curiosum.* All developed countries have complex tax structures that raise substantial amounts of revenue. Thus, the policy-relevant conclusion to be drawn from Eqn (13.6) is that even a marginal tax change can generate substantial losses in welfare, precisely because resources are shifting from an initial position in which marginal values may be far from their marginal costs. The government cannot choose to ignore the efficiency implications of minor changes in the tax structure simply because the changes are "small."

Proportional Taxes Generate No Deadweight Loss

Equations (13.6) and (13.11) indicate that the deadweight loss from distorting taxes depends fundamentally on the Slutsky substitution terms, M_{ij}. But, substitution effects can only arise from changes in relative prices that move the consumer along a given indifference curve. Thus, if all prices change in the same proportion, relative prices remain unchanged, and there can be no deadweight loss from these taxes. The compensated with-tax equilibrium is the original zero-tax equilibrium.

This can be seen directly from rewriting Eqn (13.11) as:

$$L(\vec{\mathbf{t}}) = -\frac{1}{2}\sum_{i=1}^{N} t_i \sum_{j=1}^{N} t_j M_{ij} \tag{13.13}$$

Suppose $t_j = \alpha q_j^0$, for all $j = 1,\ldots, N$, so that all prices change in the same proportion $(1 + a)$.[14] Equation (13.13) becomes:

$$L(\vec{\mathbf{t}}) = -\frac{1}{2}\sum_{i=1}^{N} t_i \sum_{j=1}^{N} \alpha \cdot q_j^0 M_{ij} = -\frac{1}{2}\sum_{i=1}^{N} t_i \alpha \sum_{j=1}^{N} q_j^0 M_{ij} \tag{13.14}$$

But, compensated demands (factor supplies), $M_i = X_i^C(\vec{\mathbf{q}}; \overline{U}^0)$, are homogeneous of degree zero in all prices. Thus $M_i[(1+\alpha)\vec{\mathbf{q}}^0; \overline{U}^0] = M_i(\vec{\mathbf{q}}^0; \overline{U}^0)$ and, from Euler's theorem on homogeneous functions, $\sum_{i=1}^{N} q_j M_{ij} = 0$, for all $\vec{\mathbf{q}}$. Hence, $L(\vec{\mathbf{t}}) = 0$.

Unfortunately, governments may not be able to use proportional taxation. With no pure profits in the system, the value of the expenditure function at the zero-tax equilibrium $M(\vec{\mathbf{q}}; \overline{U}^0)$ could well be zero, as we have been assuming. In this case, a proportional tax on all goods and factors raises no revenue because:

$$\sum_{i=1}^{N} t_i X_i^C(\vec{\mathbf{q}}; \overline{U}) = \alpha \sum_{i=1}^{N} q_i^0 X_i^C\left(\vec{\mathbf{q}}; \overline{U}^0\right)$$
$$= \alpha \sum_{i=1}^{N} q_i^0 X_i\left(\vec{\mathbf{q}}_0; \overline{U}^0\right) = 0$$

Since variable factor supplies enter the expenditure function with a negative sign, the rule "set $t_j = \alpha q_j^0$ all $j = 1, \ldots, N$" implies taxing goods and subsidizing factors, the net effect of which raises no revenue for the government.

13. Paul Samuelson discusses a number of the implications presented here in Samuelson (1986).

14. $q_j = \bar{p}_j + t_j = \bar{p}_j + \alpha q_j^0$. Hence, $q_j = \bar{p}_j + \alpha \bar{p}_j = (1+\alpha)\bar{p}_j$.

If, instead, $M(\vec{\mathbf{q}}^0; \overline{U}^0) = k, k > 0$, then a proportional tax (subsidy) on all goods and factors at rate a collects revenue equal to $\alpha \cdot k$. But, since $M(\vec{\mathbf{q}}^0; \overline{U}^0)$ includes both goods and variable factor supplies, k must be a source of lump-sum income—most likely, income from a factor in absolutely fixed supply, meaning that both its substitution and income effects are identically zero. Thus, a simpler alternative would be to tax the income from the fixed factor at rate a, a tax that cannot possibly be distorting. Henry George once proposed a tax on land rents for just this reason (George, 1914).

In conclusion, the ability to levy proportional taxes is essentially the ability to tax lump sum. For this reason, proportional taxation is hardly an interesting policy for second-best tax theory. The allocational theory of taxation ought properly concern itself only with taxes that generate distortions by changing the vector of relative prices.

Efficiency Properties of Income Taxes

About 40 years ago income taxes were held in very high regard by public sector economists. We noted in Chapter 11 that Haig–Simons income was once almost universally regarded as the ideal tax base in the ability-to-pay tradition of equity in taxation. In addition, the income tax was viewed as a highly efficient tax. The direct market effects of an income tax are on labor supply and saving behavior, and the older empirical studies found low to negligible labor and savings elasticities with respect to after-tax wages and interest rates, respectively. These studies suggested that the income tax generated very little deadweight loss.

Support for the income tax has since faded considerably. On the one hand, many neoclassical economists now prefer an expenditure tax to an income tax on dynamic efficiency and ability-to-pay grounds, as discussed in Chapter 11. On the other hand, the efficiency argument in favor of an income tax was seen to be faulty.

Nearly all the early empirical studies measured $\partial X_i/\partial q_i$, the derivative of the ordinary market supply curves with respect to changes in supply prices, whereas the relevant derivatives for efficiency loss are the Slutsky substitution effects $M_{ii} = \partial X_i^C/\partial q_i$. These two derivatives are related through the Slutsky equation, $(\partial X_i/\partial q_i) = (\partial X_i^C/\partial q_i) - X_i(\partial X_i/\partial I)$. If one observes $\partial X_i/\partial q_i = 0$, the crucial question is whether the ordinary price derivative is zero because both the substitution and the income effects are zero, or because the substitution and income effects cancel one another out. If the former is true, then indifference curves are right angled, the compensated factor supply is invariant to changes in relative prices, and taxing the factor does not produce any deadweight loss. Income from these factors is truly lump sum. If, however, the ordinary market elasticity merely reflects a canceling of income and substitution effects, the market supply curve may be vertical, but taxing the factor can, nonetheless, generate a considerable amount of deadweight loss.

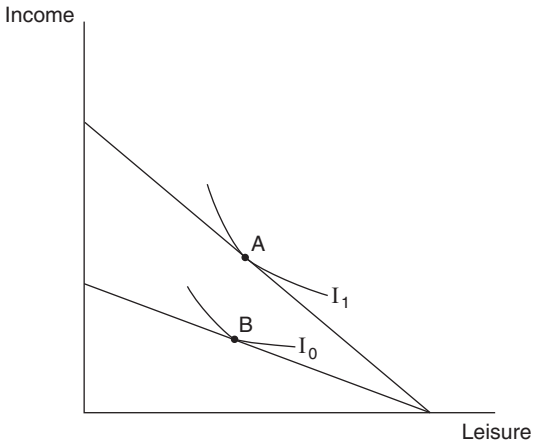

FIGURE 13.11

Unfortunately, the canceling story is quite possible for both the supply of labor and capital, since in each case the income and substitution effects work in opposite directions. Consider the case of labor, using the standard income–leisure model of neoclassical theory in which the consumer equates his marginal rate of substitution between income and leisure with the wage rate. Refer to Fig. 13.11. The slope of the budget line equals the net-of-tax wage rate. Suppose the consumer is initially in equilibrium at point A, with no tax on wages. If the government imposes a wage tax, the budget line rotates downward and the consumer reaches a new equilibrium, at point B. The question is whether B is to the right or left of A—that is, whether work effort has increased or decreased. The substitution effect of the tax says that the consumer substitutes leisure for income because the relative marginal cost of leisure (earning income) has decreased (increased). In turn, more leisure implies *less* work effort. Intuitively, marginal effort is penalized, so why work harder? The income effect says that because the price of one of the goods (income) has risen, real purchasing power has diminished. Consequently, the consumer tends to "buy" less of both goods, income and leisure. But less leisure implies *more* work effort. Intuitively, the consumer has to work harder to maintain his standard of living. Thus, the overall effect of the tax on work effort is ambiguous.

The same analysis applies to saving. A decrease in the after-tax rate of interest generates a substitution effect that favors current consumption over future consumption, and an income effect that goes against both current and future consumption. Thus, the effect on current consumption (hence saving) is ambiguous.

The empirical breakthrough regarding labor supply behavior was Jerry Hausman's (1981) paper on "Labor Supply."[15] Hausman's analysis brought a number of improvements to the estimation of labor supply, two of which were essential. He developed an estimation technique that took into account the highly nonlinear budget set that the federal personal income tax generates as a result of the personal exemption, the itemized and standard deductions, and the graduated tax rates. He also showed how to derive the indirect utility function from the estimated ordinary labor supply equation using Roy's Identity. With the indirect utility function in hand, he could solve for the compensated labor supply curve and compute appropriate measures of deadweight loss.[16]

Hausman found that the ordinary labor supply elasticity for prime-aged male heads of households was indeed very small, essentially zero, but the low elasticity was the net result of fairly substantial substitution and income effects that nearly offset one another. His point estimates of the Slutsky substitution and income elasticities were approximately 15%. The fairly high substitution elasticity led to an estimated deadweight loss of 22−54% of revenue collected, depending on income level, relative to an equal-revenue lump-sum tax. His estimates also generated a deadweight loss of 12−25% for an equal-revenue, proportional income tax. The overall estimated deadweight loss from the income tax was 28.7% of tax revenues collected.

The revised view of saving behavior came from research using neoclassical growth models in the early 1980s. Dynamic efficiency is the proper criterion for judging the effects of tax policy on saving behavior, not the static one-period effects on saving that had been the norm before 1980. Lawrence Summers (1981) was the first to suggest that taxing saving can generate large dynamic efficiency losses (Summers, 1981). He found that replacing a consumption tax with an income tax generates substantial intertemporal substitutions of consumption even if the initial effect on current consumption is quite low. His intertemporal substitution elasticities were huge, on the order of 1.2. The large intertemporal substitutions in turn generate very large dynamic efficiency losses in the long run because they induce a large reduction in the capital stock, with corresponding reductions in productivity and output. The steady-state output loss in his model was on the order of 18% of national product.

Direct versus Indirect Taxation

There is a large literature on the relative inefficiencies caused by direct versus indirect taxes, where direct taxes refer to taxes on factor supplies and indirect taxes refer to taxes on consumer goods and services. The earlier literature, which relied on static models, tended to favor direct taxation on efficiency grounds.[17] More recent literature, which employs neoclassical growth models, tends to favor indirect consumption or expenditure taxes over income taxes.[18]

In theory, Eqns (13.6) and (13.11) convey everything we need to know to settle the issue of which taxes are best. No general presumption in favor of direct taxes over indirect taxes, or vice versa, emerges from the equations. One can always postulate a set of M_{ij} (static and intertemporal) and tax rates t_k that would tip the balance one way or the other.

The issue is ultimately an empirical one. We would need to know the entire set of M_{ij} over time to resolve the issue, but our current knowledge of intertemporal Slutsky terms is virtually nil. The neoclassical growth models rely on highly simplified assumptions about the intertemporal terms, not sophisticated econometric estimates. As an empirical matter, then, statements in favor of either direct or indirect taxes must be largely conjectural, given current econometric knowledge. Furthermore, as the next section will demonstrate, the optimal pattern of taxes for raising any given amount of tax revenue is generally a mix of both direct and indirect taxes, not one or the other.

If the Government Chooses to Collect All Revenue by Imposing a Single Distorting Tax, Which Good or Factor Should It Tax?

Equation (13.11) provides the answer to this question. Consider the use of a single tax on good (factor) k versus a single tax on the good (factor) j to raise a given amount of revenue, $\overline{T}$. The loss using tax t_k is $(-)\frac{1}{2}t_k^2 M_{kk}$, assuming $t_i = 0$, for $i \neq k$. Similarly, the loss with the single tax t_j is $(-)\frac{1}{2}t_j^2 M_{jj}$. Which one dominates depends entirely on two factors: the values of M_{kk} and M_{jj} and the tax rates t_k and t_j necessary in each instance to raise the required revenue $\overline{T}$. At issue, then, is the standard empirical question: What confidence do we place in our estimates of M_{kk} and M_{jj}?

15. Hausman (1981). The estimates of elasticities and deadweight loss reported below are found on pp. 52−54, including Table 3 on p. 54, and p. 61. For an update, see Hausman and Rudd (1984).

16. We discussed Hausman's derivation of compensated consumer demand and supply functions from the ordinary estimated demand and supply functions in Chapter 9.

17. For example, see Friedman (1952) and Browning (1975). Clearly, the bias in favor of income taxes results from special assumptions in the models which, in effect, place restrictions on the values of certain M_{ij} terms.

18. Overlapping generations growth models, in particular, make a strong case for an expenditures tax on efficiency grounds. Two seminal contributions were Summers (1981) and Kotlikoff (1984).

The Issue of Tax Avoidance

People tend to favor taxes that they can avoid fairly easily, meaning taxes on goods with high price elasticities for which substitutes are readily available. But, these are precisely the taxes governments should avoid if they are concerned about deadweight loss, especially if the high price elasticities reflect large substitution effects as opposed to income effects. One immediate implication is that, on efficiency grounds alone, taxes on goods and services ought to be levied by higher rather than lower level governments in the fiscal hierarchy, that is, by the national government rather than the state governments and by state governments rather than the local governments. If a city taxes some good such as cigarettes, substitutes are readily available in the form of cigarettes sold outside the city limits. The tax artificially creates two goods in effect—city cigarettes and noncity cigarettes—that are very close substitutes.[19] A national cigarette tax is least likely to generate artificial distinctions of this type.

Single-Market Measures of Loss

Because of data limitations, empirical research is often forced to adopt partial equilibrium techniques and focus entirely on the market directly under consideration. Unfortunately, partial equilibrium measures of tax loss can be quite misleading. They would compute the loss from a tax, t_k, as $(-)\frac{1}{2}t_k^2 M_{kk}$, ignoring all cross-product terms in Eqn (13.11). This would be appropriate if t_k were the only tax, but because many goods and factors are taxed, it is not clear that the cross-product terms can be safely ignored.

One assumption commonly employed in empirical research to "justify" partial equilibrium analysis is that all cross-price elasticities are zero, but this assumption can only hold for ordinary demand (factor supply) derivatives. It cannot be imposed on the M_{ij}. Consumer theory tells us that $M_{kk} \leq 0$, and $\sum_{i=1}^{N} q_i M_{ik} = 0$, for all $k = 1,\ldots, N$. These results imply that at least one $M_{kj} \geq 0$, for $j \neq k$. In other words, if the compensated demand (supply) for one good (factor) changes in response to a relative price change, then the compensated demand (supply) for at least one other good (factor) must change as well. As always it is the substitution terms that are relevant to second-best tax questions.[20]

Despite these lessons from consumer theory, public sector economists have been willing to employ the assumption that $M_{ij} = 0$, for $i \neq j$, to get some rough indication of tax loss, even though there is no way of judging how accurate the resulting estimate is. Ignoring the cross-product terms leads to a convenient back-of-the-envelope approximation of the marginal loss per dollar of tax revenue. Write:

$$\mathrm{d}L = -t_i M_{ii} \mathrm{d}t_i \tag{13.15}$$

ignoring the cross-product terms. Assume the tax is an ad valorem percent or price tax of the form $t_i = \alpha q_i$. Therefore,

$$\mathrm{d}L = -\alpha q_i M_{ii} \mathrm{d}(\alpha q_i) \tag{13.16}$$

Because this is an approximation, calculate the marginal loss at the original equilibrium before the marginal change in the tax rate:

$$\mathrm{d}L = -\alpha q_i M_{ii} q_i \mathrm{d}\alpha \tag{13.17}$$

Multiply and divide by the original X_i to express the marginal loss in terms of the elasticity of demand for X_i, yielding:

$$\mathrm{d}L = -\alpha E_{ii} (\mathrm{d}\alpha q_i X_i) \tag{13.18}$$

The last term in the parentheses is the change in the tax revenue, $\mathrm{d}T$, given the marginal change in the tax rate a, computed at the original equilibrium. Therefore,

$$\mathrm{d}L/\mathrm{d}T = -\alpha E_{ii} \tag{13.19}$$

The marginal loss per dollar of tax revenue is approximately equal to the ad valorem tax rate times the price elasticity of demand.

Edgar Browning used this approximation to compute one of the first marginal loss estimates of the federal personal income tax, published in 1976. The relevant elasticity was the supply of labor with respect to the wage rate, which he assumed was approximately 0.2. The "average" marginal tax rate was 0.35. Therefore, his back-of-the-envelope approximation was that the federal personal income tax led to about a $0.07 increase in loss per additional dollar of revenue raised (Browning, 1976).

Browning also provided more complicated estimates based on the following considerations: (1) consumers with different incomes face different marginal tax rates, (2) exemptions and deductions in the federal income tax increase the marginal rates necessary to raise a dollar of revenue, and (3) a person's marginal tax rate is in part determined by other federal, state, and local taxes. With these additional considerations Browning was able to derive a range of estimates for $\mathrm{d}L/\mathrm{d}T$ bounded by the values $|0.07|$ and $|0.16|$. These marginal losses were considered to be comfortably low at the time.

19. City residents would also waste resources by traveling outside the city to purchase cigarettes. This waste is in addition to the standard deadweight loss.

20. Researchers will also frequently assume away all income effects, so that $\partial X_i/\partial q_j = \partial X_i^C/\partial q_j$, all $i, j = 1,\ldots, N$, in order to justify the use of ordinary demand (factor supply) derivatives in their loss measures. Given this assumption, one cannot then assume away all ordinary cross-price derivatives.

Feldstein's Estimate of Total and Marginal Deadweight Loss

Martin Feldstein more recently suggested another back-of-the-envelope calculation of the deadweight loss from income taxes that is based on the elasticity of taxable income (TI) with respect to one minus the tax rate (Feldstein, 1999). His calculations produce higher estimates of total loss per dollar of revenue than Browning-style calculations based on the labor supply elasticity, and much higher estimates of the marginal loss per dollar of additional revenue. Feldstein's calculation is based on five premises.

First, the behavioral responses to changes in personal income tax rates go far beyond changes in labor supply (and saving). They include changes in the form of compensation between taxed wages and salaries and untaxed (or more lightly taxed) fringe benefits such as contributions to pension and stock options; changes in the composition of portfolio investments; changes in itemized deductions and other expenditures that reduce TI; and changes in tax compliance. Feldstein believes that these other behavioral changes are potential sources of deadweight loss. His back-of-the-envelope calculation highlights tax avoidance through exclusions and deductions.[21]

Second, the simple utility-maximizing model behind the back-of-the-envelope calculations based on labor supply responses is

$$\max U(C, L)$$
$$\text{s.t. } C = (1-t)w(1-L)$$

where C = consumption, for which the price is assumed to be one; L = leisure; w = the wage rate; and t = the income tax rate, assumed constant for all taxpayers. Feldstein argues for an expanded model that includes exclusions (E) and deductions (D):

$$\max U(C, L, E, D)$$
$$\text{s.t. } C = (1-t)(w - wL - E - D)$$

Third, leisure, exclusions, and deductions can be considered as one composite commodity in this simple model since an income tax does not change the relative prices of L, E, and D.

Fourth, the budget constraint under a consumption (sales) tax would be $(1+\tau)C = (w - wL - E - D)$, where τ is the consumption tax rate. Therefore, a consumption tax is equivalent to the income tax if $(1+\tau)(1-t) = 1$.

Fifth, the deadweight loss (dwl) from the consumption tax equals:

$$\text{dwl} = -0.5\tau \mathrm{d}C = -0.5\tau \frac{\mathrm{d}C}{\mathrm{d}(1+\tau)}\mathrm{d}\tau \tag{13.20}$$

Refer to Fig. 13.12.

Rewriting Eqn (13.20) in elasticity form and substituting $\mathrm{d}\tau = \tau$ yields:

$$\begin{aligned}\text{dwl} &= -0.5\left(\frac{\tau}{1+\tau}\right)\left(\frac{1+\tau}{C}\right)\frac{\mathrm{d}C}{\mathrm{d}(1+\tau)}\tau C \\ &= -0.5\left(\frac{\tau}{1+\tau}\right)E_{\mathrm{C,P}}\tau C\end{aligned} \tag{13.21}$$

But, $\frac{\tau}{1+\tau} = t$ and $\tau = \frac{t}{1-t}$. Therefore,

$$\text{dwl} = -0.5t^2 E_{\mathrm{C,P}}\frac{C}{(1-t)} \tag{13.22}$$

expressed in terms of the consumption elasticity and the income tax rate.

The objective is to express Eqn (13.22) in terms of the elasticity of TI with respect to $(1-t)$. To do so, note that:

$$E_{\mathrm{C,P}} = \left(\frac{1+\tau}{C}\right)\frac{\mathrm{d}C}{\mathrm{d}(1+\tau)} = -\left(\frac{1-t}{C}\right)\frac{\mathrm{d}C}{\mathrm{d}(1-t)} \tag{13.23}$$

by replacing $(1+\tau)$ with $\frac{1}{1-t}$ and differentiating with respect to $\frac{1}{1-t}$.

The elasticity in terms of TI is

$$E_{\mathrm{TI},(1-t)} = \left(\frac{1-t}{\mathrm{TI}}\right)\frac{\mathrm{dTI}}{\mathrm{d}(1-t)} \tag{13.24}$$

From the budget constraint, the only difference between the uncompensated derivatives $\frac{\mathrm{d}C}{\mathrm{d}(1-t)}$ and $\frac{\mathrm{dTI}}{\mathrm{d}(1-t)}$ is the change in tax revenue. If the tax revenue is returned to the

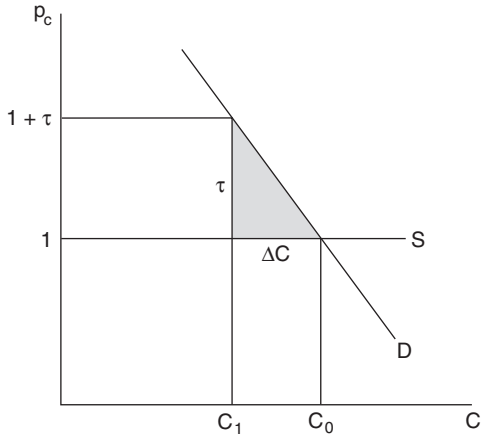

FIGURE 13.12

21. Joel Slemrod had noted in an earlier paper that the hierarchy of responses to TRA86 was (1) changes in the timing of transactions, particularly the realization of capital gains; (2) financial and accounting responses, especially the shift from nondeductible forms of debt to deductible mortgage debt; and (3) responses in "real" activities—in labor supply, saving, and investment. The real activities were a distant third in order of importance. See Slemrod (1992).

consumer, then the income-compensated derivatives are equal, or

$$\left(\frac{\mathrm{d}C}{\mathrm{d}(1-t)}\right)_{\text{comp}} = \left(\frac{\mathrm{dTI}}{\mathrm{d}(1-t)}\right)_{\text{comp}} \tag{13.25}$$

Compensating to hold income constant is not the same as compensating to hold utility constant. Nonetheless, Feldstein is following Harberger here. Harberger recommended using income rather than utility-compensated elasticities in applied work (Harberger, 1971).

Combining Eqn (13.25) with the definitions of the two elasticities, Eqns (13.23) and (13.24), implies:

$$\begin{aligned} -CE_{\mathrm{C,P}} &= (1-t)\left(\frac{\mathrm{d}C}{\mathrm{d}(1-t)}\right)_{\text{comp}} \\ &= (1-t)\left(\frac{\mathrm{dTI}}{\mathrm{d}(1-t)}\right)_{\text{comp}} = \mathrm{TI}\cdot E_{\mathrm{TI},(1-t)} \end{aligned} \tag{13.26}$$

Therefore, from Eqn (13.22), deadweight loss in terms of TI is

$$\mathrm{dwl} = +0.5t^2 E_{\mathrm{TI},(1-t)}\frac{\mathrm{TI}}{(1-t)} \tag{13.27}$$

This is the back-of-the-envelope formula Feldstein recommends for calculating the deadweight loss and comparing it with the total tax revenue. Further, the difference in the deadweight loss for two different tax rates is the estimate of the marginal deadweight loss from replacing one tax rate with the other. The marginal deadweight loss can then be compared with the change in TI from the tax change.

In previous work analyzing the Tax Reform Act of 1986 (TRA86), Feldstein estimated that the average $E_{\mathrm{TI},\ (1\ -\ t)} = 1.04$ across all taxpayers (Feldstein, 1995). He believes this can be considered an income-compensated elasticity because TRA86 was designed to be revenue neutral across households. This elasticity yields the following total and marginal deadweight loss estimates:

1. In 1994, the deadweight loss of the personal income tax was 32.2% of the total tax revenues.
2. Raising all marginal tax rates in 1994 by 10% would generate a deadweight loss of $43 billion.

Tax revenues, however, would only increase by $26 billion, and only by $21 billion counting the reduction in payroll taxes caused by the reduction in labor supply. Using the $21 billion produces an incremental deadweight loss of $2.06 per dollar of additional tax revenue. This estimate far exceeds the usual estimates in the previous literature. Nonetheless, Feldstein cautions that tax increases that primarily affect the highest income taxpayers, such as the 1993 tax reform, are likely to generate even larger increases in deadweight loss per dollar of additional tax revenue. His analysis of TRA86 suggested that $E_{\mathrm{TI},\ (1\ -\ t)}$ was on the order of 3 for the highest income taxpayers. Therefore, increasing their tax rates generates very high deadweight losses and very little additional tax revenue, leading to extremely high incremental deadweight losses per dollar of revenue collected.

The accuracy of Feldstein's suggested back-of-the-envelope calculation is difficult to judge. First, income-compensated elasticities are not the same as utility-compensated elasticities, and the latter can only be determined by specifying how specific elements of E and D enter the utility function.[22] Second, the elasticity of consumption with respect to income tax rates could be quite low, especially among high-income taxpayers. One wonders how accurate Feldstein's calculation would be relative to a richer intertemporal model that allows for saving. The equivalence of the income and consumption tax is not as simple as in the static model. The income tax would have to allow for deductions of all income from capital to establish equivalence. Second, income shifting between untaxed and taxed items may not have much effect on consumption and saving, just on the form of saving. If the principal changes are to households' budget constraints and not directly to their utilities, then the total and (especially) the marginal deadweight losses from the personal income tax may be far less than Feldstein's simple calculations suggest. One's intuition is that responses of the real activities—of labor supply, saving, and investment—are the major sources of deadweight loss arising from an income tax. Nonetheless, the other behavioral responses that Feldstein mentions are important reactions to income taxes and they may well increase the deadweight loss from income taxes to the extent that Feldstein's measure suggest.

Gruber and Saez on the Elasticity on TI

Feldstein's analysis prompted a number of studies on the elasticity of TI. Particularly noteworthy is the study by John Gruber and Emmanuel Saez that used a panel of tax returns from the National Bureau of Economic Research (NBER) to estimate the responses to all the state and federal changes in tax rates from 1979 to 1990. Feldstein's empirical estimates were based solely on the large federal tax reform, TRA86. Gruber and Saez's richer data set allowed them to consider the reactions of different income groups. They report results for three income ranges: $10,000−$50,000, $50,000−$100,000, and greater than $100,000. This is an important addition given the graduated federal tax rates. They also developed an econometric approach that allowed

22. An excellent analysis of the potential pitfalls of substituting income-compensated for utility-compensated elasticities in loss measures is Richter (1977).

them to estimate the substitution and income effects of the tax rate changes. Finally, they computed elasticity responses of "broad income," essentially the total income reported on federal tax returns (excluding capital gains, which are subject to different rates) and TI, essentially broad income less exclusions and deductions (and the personal exemptions available to all taxpayers).

Their principal results were the following:

1. The average elasticity of TI with respect to changes in tax rates across all income groups was 0.4, well below Feldstein's estimate of 1.04. The elasticity of broad income was only about one-third as large. The 0.4 estimate is midway in the range of estimates of studies of the elasticity of TI, most of which are between zero and 0.8.
2. The average elasticity of TI masks considerable variation across the three income categories: 0.18 in the $10,000−$50,000 range; 0.11 in the $50,000−$100,00 range, and 0.57 for taxpayers with TI greater than $100,000. The elasticities for broad income are lower for all groups. The point estimates are negative for the two lower groups and 0.17 for the highest group. A caveat is that none of the estimates for taxable or broad income are statistically significant. Nonetheless, Gruber and Saez believe that the estimates on the highest income group suggest that they respond to changes in tax rates primarily by varying the deductions and exclusions that they take. The elasticity of broad income will necessarily be lower than the estimates for TI simply because broad income is a larger number. But they believe this accounts for only two-fifths of the difference between the 0.57 and 0.17 estimates for the highest group. The remaining three-fifths is likely due to the variation in deductions and exclusions. Gruber and Saez speculate that Feldstein obtained a much higher elasticity estimate because he focused on TRA86, in which most of the big tax rate changes affected the higher income taxpayers.
3. Estimated income effects of broad income in response to the tax rate changes are very low, indicating that there is little difference between the actual and compensated elasticities in terms of the effects on labor supply. This is much different from the original Hausman study of labor supply elasticities reported earlier in the chapter.
4. These last two results imply that redistribution through the tax system should consist of a large grant given to all taxpayers within the phase-out region subject to high tax rates, and then relatively constant or even declining marginal tax rates in the higher income range. Given the low actual and compensated elasticities in the lower two ranges, high tax rates in the phase-out range will not generate much of a labor supply response or much deadweight loss. In contrast, keeping marginal tax rates relatively low on the highest income groups will reduce the overall deadweight loss of an income tax.
5. Taxpayers respond more to changes in state income tax rates than to changes in federal tax rates.[23]

Efficiency Cost of the Personal Income Tax

No consensus exists on the total or marginal deadweight loss from the federal personal income tax. As noted earlier, Hausman estimated the total loss per dollar of revenue collected at 28.7% looking only at the labor supply response. Feldstein's estimate considering other behavioral responses of 32.2% is in the same ballpark (Gruber and Saez did not provide a deadweight loss estimate of the income tax). The accuracy of these numbers is subject to debate, however. Thomas MaCurdy has shown that Hausman's estimating procedure can produce an upward bias in the estimate of the substitution effect, suggesting that the deadweight loss was overstated (MaCurdy, 1992). Other researchers have noted that estimates of female labor supply elasticities are typically much higher than those of males, suggesting that the deadweight loss has increased as more women entered the labor market. A recent study of female labor supply responses in England by Richard Blundell et al., however, found compensated wage elasticities ranging from 0.14 to 0.44, much smaller than in most previous studies.[24] Finally, the elasticity of TI estimated by Gruber and Saez, and most other studies, is much lower than Feldstein's estimate, implying a lower efficiency loss from income taxes. Therefore, to claim that a consensus has arisen of about 30% of the revenue collected may not be warranted.

Whatever the estimate of deadweight loss, it should be increased by the costs of complying with income tax, the second potential source of inefficiency. Marsha Blumenthal and Joe Slemrod have estimated the compliance costs at about 5−7% of the revenue collected (Blumenthal and Slemrod, 1992). The administrative costs of collecting the revenue, the third potential source of inefficiency, are minuscule and can be ignored. There is no controversy on this point.

The estimates of the marginal deadweight loss per additional dollar of revenue collected are all over the lot. The weight of the evidence suggests that the marginal loss is larger than the average loss, and perhaps much larger, but few studies have estimated marginal losses above $1 per

23. Gruber and Saez (2002). The results reported above are in Tables 4 (overall elasticities), 5 (substitution and income effects), 8 (federal versus state tax rate responses), and 9 (estimates for the three income groups).
24. Blundell et al. (1998). The elasticity estimates are from Table IV, p. 846. See, also, Blundell and MaCurdy's review of the labor supply literature in Blundell and MaCurdy (1999).

dollar of additional revenue. Feldstein's $2.06 estimate is well above the typical estimate in the literature.

THE OPTIMAL PATTERN OF COMMODITY TAXES

The second main question of distorting taxation is that of optimality. Suppose the government has to raise a given amount of tax revenue $\overline{T}$, subject to the constraint that it must use distorting unit taxes paid by the consumer. If the government is free to tax or subsidize any good (or factor), what pattern of taxes raises the required revenue in such a way as to minimize deadweight loss in the economy? This is commonly referred to as the optimal commodity tax problem, with the understanding that "commodities" include both goods and factors.

Having defined the appropriate loss function for a one-consumer, linear technology economy, the optimal commodity tax problem is a straightforward, constrained, loss-minimization problem of the form[25]:

$$\min_{(t_k)} L(\vec{\mathbf{t}})$$
$$\text{s.t.} \quad \sum_{i=1}^{N} t_i X_i^{\mathrm{C}} = \overline{T}$$

Two assumptions are necessary to ensure that the first-order conditions yield interesting results. First, we will continue to assume that the value of the consumer's expenditure function in the pretax equilibrium is zero; that is, $M(\vec{\mathbf{p}}; \overline{U}^0) = 0$. If, to the contrary, $M(\vec{\mathbf{p}}; \overline{U}^0) = k$, the loss-minimizing strategy is to tax the lump-sum income k at rate a such that $\overline{T} = \alpha k$, thereby avoiding any loss at all. If $k < \overline{T}$, the loss-minimizing strategy is to tax away the entire k and then using distorting taxes to collect revenue equal to $(\overline{T} - k)$. Because the problem can always be redefined in this way, it is convenient to assume $k = 0$ at the outset so that lump-sum income taxation is impossible.

Second, with zero lump-sum income, both demand and supply are homogeneous of degree zero in prices $\vec{\mathbf{q}}$ and $\vec{\mathbf{p}}$. Hence, we are permitted two separate normalizations, one for $\vec{\mathbf{q}}$ and one for $\vec{\mathbf{p}}$. For convenience, normalize both on the same good, the first. Therefore, set $q_1 \equiv p_1 \equiv 1$, which in turn implies $t_1 = 0$, or that the government does not tax the first good. These normalizations also remove the uninteresting possibility of proportional taxation, which is equivalent to lump-sum taxation and would entail no deadweight loss. Proportional taxes also raise no revenue given the assumption of zero lump-sum income and constant producer prices. With good one untaxed, however, any set of tax rates on the remaining $(N - 1)$ goods and factors necessarily changes the vector of *relative* prices and generates loss.[26]

Given these two assumptions, define the Lagrangian[27] for the optimal tax problem as:

$$\min_{(t_k)} L = L(\vec{\mathbf{t}}) - \lambda\left(\sum_{i=1}^{N} t_i X_i^{\mathrm{C}} - \overline{T}\right) = M\left(\vec{\mathbf{q}}; \overline{U}^0\right) - \sum_{i=1}^{N} t_i X_i^{\mathrm{C}} - \lambda\left(\sum_{i=1}^{N} t_i X_i^{\mathrm{C}} - \overline{T}\right)$$

with $t_1 \equiv 0$. The first-order conditions are

$$\frac{\partial L}{\partial t_k} = X_k^{\mathrm{C}} - X_k^{\mathrm{C}} - \sum_{i=1}^{N} t_i \frac{\partial X_i^{\mathrm{C}}}{\partial q_k} - \lambda\left(X_k^{\mathrm{C}} + \sum_{i=1}^{N} t_i \frac{\partial X_i^{\mathrm{C}}}{\partial q_k}\right) = 0, \quad \text{for } k = 2, \ldots, N \tag{13.28}$$

(recall that $\partial q_k = \partial t_k$ with linear technology). Rearranging terms:

$$-(1+\lambda)\left(\sum_{i=1}^{N} t_i \frac{\partial X_i^{\mathrm{C}}}{\partial q_k}\right) - \lambda X_k^{\mathrm{C}} = 0, \quad \text{for } k = 2, \ldots, N, \text{ or} \tag{13.29}$$

$$\frac{\sum_{i=1}^{N} t_i \frac{\partial X_i^{\mathrm{C}}}{\partial q_k}}{X_k^{\mathrm{C}}} = \frac{-\lambda}{1+\lambda} \quad k = 2, \ldots, N \tag{13.30}$$

Also,

$$\sum_{i=1}^{N} t_i X_i^{\mathrm{C}} = \overline{T} \tag{13.31}$$

Notice that the right-hand side (RHS) of Eqn (13.30) is independent of k. Furthermore, since $\partial X_i^{\mathrm{C}}/\partial q_k = M_{ik} = M_{ki} = \partial X_k^{\mathrm{C}}/\partial q_i$ from the symmetry of the Slutsky substitution terms, Eqn (13.30) can be rewritten as:

$$\frac{\sum_{i=1}^{N} t_i M_{ki}}{X_k^{\mathrm{C}}} = C \quad k = 2, \ldots, N \tag{13.32}$$

The numerator, $\sum_{i=1}^{N} t_i(\Delta X_k^{\mathrm{C}}/\Delta q_i) = \sum_{i=1}^{N} t_i(\Delta X_k^{\mathrm{C}}/\Delta t_i)$ approximates the total change in X_k in response to marginal changes in all the taxes, 2, …, N. Hence, the left-hand side (LHS), $\Delta X_k^{\mathrm{C}}/X_k^{\mathrm{C}}$, is the percentage change in X_k^{C}, in response to the tax package. The first-order conditions, then, require a set of taxes that produce equal percentage

25. The optimal tax problem can also be modeled using the consumer's utility function as the objective function, in which the goal is to maximize utility subject to a revenue constraint and distorting taxation. The resulting tax rules are identical upon using the Slutsky equation to substitute the compensated demand (and factor supply) derivatives for the ordinary derivatives.

26. We assume further that the revenue requirement $\overline{T}$ is feasible.

27. Whether the tax revenue summations go from 1,…, N or 2,…, N is immaterial as $t_1 \equiv 0$.

changes in the compensated demands and supplies for all goods and factors.[28,29]

Notice that Eqn (13.32) describes percentage changes in terms of quantities and not the tax rates themselves. Unfortunately, the pattern of tax rates cannot be described by an equivalently simple rule. Their general pattern is clear, however: goods (factors) whose compensated demands (supplies) are relatively inelastic should be subjected to relatively higher rates of taxation. This is the only way the "equal percentage change" rule can possibly be satisfied.

A rough intuitive explanation of the rule can be obtained by considering the optimal tax problem as one of minimizing the sum of the deadweight loss triangles in each market (this ignores the cross-substitution terms M_{ij}, $j \neq i$, and their corresponding loss rectangles depicted in Fig. 13.10).[30]

If the compensated demand for one good is highly inelastic and the compensated demand for another highly elastic, most of the required revenue should be raised from the inelastic good. Its quantity demanded does not change much even with a relatively high tax rate. Consequently, it raises a relatively large amount of revenue with a relatively small deadweight loss triangle. Conversely, placing an equal tax rate on a good with a relatively elastic demand causes a larger change in quantity demanded. Hence, the revenue collected is smaller and the deadweight loss triangle larger. Per dollar of revenue then, it pays to tax the relatively inelastic goods (factors). In the limit, if one good (factor) has a perfectly inelastic compensated demand (supply), it should be used to collect all the revenue. There can be no deadweight loss, and the percentage change in output is equalized across all goods at a value equal to zero.

Policy Implications of the Optimal Tax Rule

The equal percentage change rule is a deceptively simple representation of the optimal tax equilibrium. Computing the actual tax rates involves solving N first-order conditions for λ and the t_k, $k = 2,\ldots, N$, a prodigiously complex task, especially given limited econometric knowledge of the crucial Slutsky substitution terms. Furthermore, all goods and factors (except the untaxed numeraire) are either taxed or subsidized at the optimum, in general. Thus, it is extremely unlikely that any government could ever even approximate the *optimal* pattern of tax rates. Nonetheless, the equal percentage charge rule yields a number of useful qualitative insights for tax policy.

Broad-Based Taxation

The optimal commodity tax rule (Eqn (13.32)) offers a strong presumption against broad-based taxes such as general sales or general income taxes, which tax a broad range of goods or factors at a single rate. Additional restrictions on preferences are clearly required to generate the result that $t_k = aq_k$ with k defined over two or more goods (or factors).

Public sector economists have been particularly interested in the restrictions on preferences required for uniform taxation, in which all goods are taxed at the same proportional rate, the broadest general sales tax. Understanding the conditions for uniform taxation is especially compelling because of a theorem we will prove in Chapter 16 that the efficiency implications of an equal proportional tax on all the goods can be duplicated by replacing it with a proportional tax on all factors. Hence, if a uniform sales tax is optimal, it need not be used. A uniform income tax is also optimal. Income taxes are generally preferred to sales taxes on equity grounds because they are easier to tailor to the personal circumstances of individuals and families.

The public sector literature contains a number of sufficient conditions for uniform taxation in a model with $(N-1)$ goods and labor as the untaxed numeraire. The first results appeared in the early 1970s. In 1995, Timothy Besley and Ian Jewitt were finally able to establish the necessary and sufficient conditions for uniform taxation in terms of a concept called the wage-compensated labor supply.[31] A sketch of the proof follows.

Distinguishing between the goods and labor is useful. Define $\vec{\mathbf{X}}$ as the N-vector of goods, with element X_k; L,

28. If all the taxed goods and factors, $k = 2,\ldots N$, undergo equal percentage changes, then the first untaxed good also undergoes the same percentage change, although the base for computing the percentage change differs.

29. The equal percentage change interpretation applies, strictly speaking, only to marginal changes in each of the tax rates from the no-tax position, that is, to a marginal revenue package. For discrete tax changes, the rule implies that there must be an equal percentage change in quantity demanded in response to a further infinitesimal proportional change in all the tax rates from their optimum values. This interpretation is necessary because the compensated demand curves in the discrete case are all evaluated at the gross-of-tax consumer prices existing at the optimum when solving for the optimal pattern of the t_k. If all the M_{ik} are constant in the relevant range, however, then the rule needs no modification for the discrete case. This follows because: $\Delta X_k^C = \sum_{i=1}^{N} \int_0^{t_i} M_{ik} \mathrm{d}q_i = \sum_{i=1}^{N} \int_0^{t_i} M_{ki} \mathrm{d}q_i = \sum_{i=1}^{N} M_{ki} \int_0^{t_i} \mathrm{d}q_i = \sum_{i=1}^{N} t_i M_{ki} = \sum_{i=1}^{N} t_i M_{ik}$ when the M_{ik} are constant (and $t_1 \equiv 0$).

30. See Baumol and Bradford (1970), on this point. Their article offers an excellent intuitive feel for the optimal tax problem and the properties of its solutions.

31. Their model is identical to our simple model with the exception that it allows for general technology with CRS production. As we will see in Chapter 14, however, the first-order conditions for optimal taxation with general technology continue to be Eqn (13.32) as long as there are no pure profits, which is true under CRS production (and perfect competition). See Besley and Jewitt (1995).

as labor, the untaxed numeraire; $\vec{\mathbf{q}}$, as the N-vector of goods prices with element q_k; and w, as the wage.

Necessary Conditions

If taxes are optimal, then Eqn (13.32) holds for all $k = 1,\ldots,N$:

$$\sum_{i=1}^{N} t_i M_{ik} = -\theta M_k \tag{13.33}$$

Suppose also that taxes are uniform, so that $t_i = \rho q_i$, $i = 1, \ldots, N$. Therefore,

$$\sum_{i=1}^{N} \rho q_i M_{ik} = -\theta M_k, \quad \text{or} \tag{13.34}$$

$$\sum_{i=1}^{N} q_i M_{ik} = (-\theta/\rho) M_k \tag{13.35}$$

Homogeneity of the compensated demands and supplies implies:

$$\sum_{i=1}^{N} q_i M_{ki} + w M_{kw} = 0 \tag{13.36}$$

Therefore,

$$w M_{kw} = (\theta/\rho) M_k, \quad \text{or} \tag{13.37}$$

$$M_{kw} = [\theta/(\rho w)] M_k = \alpha M_k \quad k = 1,\ldots,N \tag{13.38}$$

where a is a scalar. The necessary conditions for uniform taxation are that the derivatives of the compensated demand for each good with respect to the wage are proportional to the compensated demand for the good.

Sufficient Condition

The sufficient condition relies on the property that the producer and consumer prices of all the goods must be collinear for Eqns (13.32) and (13.38) to hold simultaneously. That is, $p_i = kq_i$, for $i = 1, \ldots, N$. But, $t_i = q_i - p_i$. Therefore, the collinearity condition implies uniform taxes (see their article for the details surrounding the collinearity condition on the prices).

Consider, next, the wage-compensated labor supply. *Wage compensated* means that the wage adjusts to maintain utility at a given level. That is, $w = w(q, \overline{U})$ such that $V(q, w(q, \overline{U})) = \overline{U}$, where V is the indirect utility function. The wage-compensated labor supply is the derivative of the expenditure function with $w = w(q, \overline{U})$:

$$L(q, \overline{U}) = -\partial M\big(q, w(q, \overline{U}), \overline{U}\big)/\partial w \tag{13.39}$$

To obtain the necessary and sufficient conditions in terms of the wage-compensated labor supply, differentiate $L(q, \overline{U})$ with respect to q_k:

$$\partial L/\partial q_k = -\partial M^2/\partial w \partial q_k + \partial^2 M/\partial w^2 \cdot \partial w/\partial q_k \tag{13.40}$$

$$= -\partial M^2/\partial w \partial q_k + \partial^2 M/\partial w^2 \cdot (\partial M/\partial q_k/\partial M/\partial w) \tag{13.41}$$

$$= -\partial M^2/\partial w \partial q_k + \big(\partial^2 M/\partial w^2/\partial M/\partial w\big) \cdot \partial M/\partial q_k, \quad k = 1,\ldots,N \tag{13.42}$$

The RHS of Eqn (13.42) has the same form as Eqn (13.38). Therefore, the necessary and sufficient conditions for uniform taxation are that the RHS of Eqn (13.42) equal zero, or $\partial L(q, \overline{U})/\partial q_k = 0$, all $k = 1, \ldots, N$. The wage-compensated labor supply curve must be independent of all commodity prices for uniform taxation to be optimal.

An immediate implication of the Besley–Jewitt theorem is that uniform taxation is optimal if[32]:

$$\partial\big(\partial M/\partial q_i/\partial M/\partial q_j\big)/\mathrm{d}w = 0 \text{ all } i,j = 1,\ldots,N \tag{13.43}$$

or

$$\partial\left(\frac{X_i^C}{X_j^C}\right)\Big/\partial w = 0 \tag{13.44}$$

The ratio of the compensated demands is independent of the wage at the optimum. The literature has developed a number of sufficient conditions for the optimality of uniform taxation based on the separability of either the expenditure function or the utility function, all of which imply condition (13.44). For example, an expenditure function of the form $M(q, w, U) = F(g(q, U), w, U)$ satisfies Eqn (13.44).[33,34]

32. $\partial(\partial M/\partial q_i/\partial M/\partial q_j)/w = (\partial M/\partial q_j \cdot \partial^2 M/\partial q_i \partial w - \partial M/\partial q_i \cdot \partial^2 M/\partial q_j \partial w)/(\partial M/\partial q_j)^2$. But, with uniform optimal taxation, $\partial^2 M/\partial q_i \partial = \alpha \partial M/\partial q_i$ and $\partial^2 M \partial q_j \partial w = \alpha \partial M/\partial q_j$, so that the numerator is zero.

33. See Besley and Jewitt, *op. cit.*, for a complete analysis of why separability of the expenditure function is sufficient but not necessary for uniform taxation to be optimal.

34. Atkinson and Stiglitz were the first to derive sufficient conditions for uniform taxation in the simple model with $N - 1$ goods and labor the untaxed numeraire. Atkinson and Stiglitz prove that uniform taxation is optimal if either (1) labor is in absolutely fixed supply (see pp. 319–320) or (2) preferences are homothetic. They also proved that if preferences have an additive representation (i.e., $U(X_1\ldots, X_{n-1}\ L) = {}_{g1}(X_1) + \ldots + g_{n-1}(X_{n-1}) + g_n(L)$), then tax rates are inversely proportional to each commodity's income elasticity of demand. This implies uniform taxation if preferences are additive in logarithms with equal coefficients, in which case all income elasticities equal 1. They also proved that if preferences have an additive representation, and the marginal disutility of labor is constant (i.e., $\partial g_n/\partial L = k$), then the optimal tax rates are inversely proportional to each commodity's own price elasticity of demand (refer to the discussion of the IER, below). See Atkinson and Stiglitz (1972).

The Exemption of "Necessities"

The optimal commodity tax rule also gives a strong presumption against the common practice of exempting necessities such as food and clothing from sales tax bases. If anything, these items can be expected to have relatively low substitution effects (along with their income elasticities being less than 1). Therefore, by the efficiency criterion, they should be taxed at *higher* than average rates, not exempted from taxation. But, governments exempt these items anyway for equity reasons, in an attempt to make sales taxes somewhat less regressive. For example, 26 states in the United States exempt food purchased for home consumption from their sales taxes.

Analysis of optimal commodity taxation within the context of a many-consumer economy, the subject of Chapter 14, can reconcile the equity−efficiency trade-off, but only in principle. Many-person tax rules are extremely difficult to apply. Nonetheless, it is clear that many governments have been swayed more by equity than by efficiency arguments in designing their sales taxes. This is often the case whenever equity and efficiency goals conflict. Favoring equity over efficiency considerations is not peculiar to tax policy.

Percentage Charge Rules for Ordinary Demand (Factor Supply) Relationships

Some additional qualitative policy information can be obtained by rewriting Eqn (13.32) in terms of the ordinary price and income derivatives by means of the Slutsky equation:

$$\frac{\partial X_k}{\partial q_i} = M_{ki} - X_i \frac{\partial X_k}{\partial I} \tag{13.45}$$

or

$$M_{ki} = \frac{\partial X_k}{\partial q_i} + X_i \frac{\partial X_k}{\partial I} \tag{13.46}$$

Substituting Eqn (13.46) into (13.32) yields:

$$\frac{\sum_{i=1}^{N} t_i \left(\frac{\partial X_k}{\partial q_i} + X_i \frac{\partial X_k}{\partial I} \right)}{X_k} = C \quad k = 2, \ldots, N \tag{13.47}$$

Rearranging terms:

$$\frac{\sum_{i=1}^{N} t_i \frac{\partial X_k}{\partial q_i}}{X_k} = C - \sum_{i=1}^{N} t_i X_i \frac{\frac{\partial X_k}{\partial I}}{X_k} \quad k = 2, \ldots, N \tag{13.48}$$

Multiplying and dividing the second term on the RHS of Eqn (13.48) by I yields:

$$\frac{\sum_{i=1}^{N} t_i \frac{\partial X_k}{\partial q_i}}{X_k} = C - \frac{\sum_{i=1}^{N} t_i X_i}{I} (E_{k,I}) \quad k = 2, \ldots, N \tag{13.49}$$

where $E_{k,\,I}$ is the income elasticity for good k.

Assuming that $\partial X_k/\partial q_i$ are constant in the relevant range, the LHS of Eqn (13.49) gives the percentage change in the ordinary demand (supply) of the kth good (factor). Notice that these percentage changes are not equal. Goods with higher income elasticities should change by the greater amount (in absolute value; C is presumably negative for goods). The intuition is to exploit income effects, since they do not contribute to deadweight loss. Of course, the optimal tax rates are no more easily solved by Eqn (13.49) than by Eqn (13.32) (including the revenue constraint in each instance). At the same time, common sense often suggests which goods tend to have relatively high income elasticities. Notice, for example, that Eqn (13.49) gives a partial efficiency justification for taxing necessities lightly.

The Inverse Elasticity Rule

An approximation to the equal percentage change rule that is often used in policy analysis is the inverse elasticity rule (IER). The IER says that tax rates should be increased in inverse proportion to a good's (factor's) price elasticity of demand.[35] The basis for this interpretation of Eqn (13.32) is as follows.

Suppose, as an approximation, that all income effects are ignored as being empirically unimportant and, further, that all cross-price derivatives are set equal to zero on the grounds that their own price effects dominate the cross-price effects. With these two assumptions, Eqn (13.32) reduces to:

$$\frac{t_k M_{kk}}{X_k} = C \quad k = 2, \ldots, N \tag{13.50}$$

where:

M_{kk} = the own price derivative for *both* compensated and ordinary demand (supply) curves, since there are no income effects.

Multiplying and dividing the LHS of Eqn (13.50) by q_k yields:

$$\left(\frac{t_k}{q_k} \right) \cdot \left(\frac{\partial X_k}{\partial q_k} \cdot \frac{q_k}{X_k} \right) = C \quad k = 2, \ldots, N \tag{13.51}$$

Alternatively,

$$\left(\frac{t_k}{q_k} \right) = \frac{C}{E_{kk}} \quad k = 2, \ldots, N \tag{13.52}$$

where:

E_{kk} = the own-price elasticity of demand.

35. Kahn (1970), contains a discussion of the IER in the context of price discrimination. Kahn's analysis reflects the importance of the IER in the industrial organization literature. As we shall discover in Chapter 23, second-best pricing rules for multiproduct decreasing-cost industries with profit constraints are virtually identical to the optimal tax rule.

The tax rate as a percentage of the gross-of-tax price, q, should be inversely proportional to the own-price elasticity of demand (supply) for each good (factor), hence the IER.

The IER is an intuitively appealing interpretation of the optimal tax rules, especially if one thinks in terms of minimizing deadweight loss triangles, but the assumptions supporting this interpretation are heroic, to say the least. As noted in the preceding section on marginal loss, if all income effects are zero, then ordinary price derivatives must follow the same laws as compensated price derivatives, in particular the homogeneity result that $\sum_{i=1}^{N} q_i M_{ik} = 0$, all $k = 1,\ldots, N$. But, this implies $M_{ki} \neq 0$ for some i, $i \neq k$ (because $M_{ii} < 0$). One legitimate possibility is to assume that $M_{ki} = 0$, $i \neq k$, for all *taxed* goods (factors). This implies that all cross-price effects occur with respect to the untaxed numeraire; that is, $M_{ki} \neq 0$, all $k = 2,\ldots, N$. In particular, with $q_i \equiv 1$ (the numeraire),

$$M_{ki} = -q_k M_{kk} \quad k = 2, \ldots, N \tag{13.53}$$

Unfortunately, Eqn (13.53) can hardly be expected to be true. Thus, the IER may not be very useful even as a rough guideline to the policy maker.

A more sensible alternative for policy analysis may be to select one or two M_{ki}, $i \neq k$, that are likely to be nonzero; place reasonable values on them that satisfy the homogeneity condition $\sum_{i=1}^{N} q_i M_{ik} = 0$; and apply a simplified version of the equal percentage change rule, Eqn (13.32). It will still have nearly an inverse elasticity interpretation. For example, suppose one assumes that only M_{kk} and M_{kj} are nonzero when evaluating the first-order condition for t_k. The kth relation in Eqn (13.32) becomes:

$$\frac{t_k M_{kk} + t_j M_{kj}}{X_k} = C \tag{13.54}$$

Multiplying and dividing the two terms on the LHS by q_k and q_j, respectively, yields:

$$\frac{t_k}{q_k} \cdot E_{kk} + \frac{t_j}{q_j} E_{kj} = C \tag{13.55}$$

Rearranging terms,

$$\frac{t_k}{q_k} E_{kk} = C - \frac{t_j}{q_j} E_{kj} \tag{13.56}$$

and

$$\left(\frac{t_k}{q_k}\right) = \frac{1}{E_{kk}} \left[C - \left(\frac{t_j}{q_j}\right) E_{kj} \right] = \left[C/E_{kk} + t_j/q_j \right] \tag{13.57}$$

with $E_{kj} = -E_{kk}$. In this form, the IER says that the percentage tax on good (factor) k is inversely related to its own-price elasticity corrected by a term equal to the percentage tax (at the optimum) on the other good.

This simplification at least avoids having to impose patently unrealistic assumptions on the compensated cross-price elasticities. Also, the resulting simultaneous system of equations would not be much more difficult to solve than the standard IER applied to all goods and factors.[36]

SUBSTITUTIONS AMONG TAXES: IMPLICATIONS FOR WELFARE LOSS

The third main question of distorting taxation is that of tax reform: What is the implication on social welfare of substituting one set of taxes for another while holding revenue constant? This tax substitution experiment is perhaps the most compelling of all second-best exercises within the pure allocational theory of taxation, if only because governments occasionally engage in such tax substitutions. We continue to assume a one-consumer-equivalent economy with linear technology.

As long as the tax changes are "small," the expressions for marginal loss and total tax revenue are all that are needed to determine the efficiency implications for any given equal-revenue substitution among taxes. Begin with the total differential of deadweight loss, Eqn (13.11), with respect to all the taxes:

$$\mathrm{d}L = -\sum_{k=1}^{N} \sum_{i=1}^{N} t_i M_{ik} \mathrm{d}t_k \tag{13.58}$$

(there is no need to assume an untaxed good in this exercise). Equation (13.58) is an appropriate measure of marginal loss for any given change in the vector of tax rates. The reason why it is a substitution can always be viewed as a multistep series of individual loss experiments, in which one tax is reduced and the revenue returned to the government lump sum, after which a second tax is imposed, with its revenue returned to the consumer lump sum, and so on, for any number of tax changes. Because $M_{ik} = M_{ki}$, the order of substitution is irrelevant.

Next, add the values of $\mathrm{d}t_k$ that hold revenue constant. These can be determined by totally differentiating the tax revenue equation:

$$\mathrm{d}T = \sum_{k=1}^{N} \left(M_k + \sum_{i=1}^{N} t_i M_{ik} \right) \mathrm{d}t_k \tag{13.59}$$

Setting $\mathrm{d}T = 0$, Eqn (13.59) determines all possible tax substitutions that keep the revenue unchanged. Once the

36. The IER has figured prominently in public hearings concerned with setting prices in the regulated industries. One common example is the US postal service, which from 1974 to 1980 used the IER to justify its policy of covering its cost increases primarily by increasing rates on first-class mail (relatively inelastic demands) rather than on the other classes of mail such as parcel post (relatively elastic demands). We will return to the postal service example in Chapter 23.

appropriate values for dt_k have been determined from Eqn (13.59), they can be substituted back into Eqn (13.58) to determine the resulting change in deadweight loss.

When only two taxes change, Eqn (13.59) describes the exact relationship between the two changes necessary to hold revenue constant. Suppose, for example, t_j and t_k are to be changed, $dt_i = 0$, for $i \neq j, k$. From Eqn (13.59),

$$dT = 0 = \left(M_k + \sum_{i=1}^{N} t_i M_{ik}\right) dt_k + \left(M_j + \sum_{i=1}^{N} t_i M_{ij}\right) dt_j \tag{13.60}$$

or

$$\frac{dt_k}{dt_j} = -\frac{\frac{\partial T}{\partial t_j}}{\frac{\partial T}{\partial t_k}} \tag{13.61}$$

As expected, the two rates must change in direct ratio to the marginal changes in tax revenue with respect to each of the taxes. Presumably, one tax is increased and the other is decreased. Notice also that the relevant marginal revenue changes are the changes at the compensated equilibria, not the actual equilibria. This is consistent with the definition of loss in terms of compensated equilibria.

To complete the analysis, the marginal loss with respect to changes in t_j and t_k is

$$dL = -\left(\sum_{i=1}^{N} t_i M_{ik} dt_k + \sum_{i=1}^{N} t_i M_{ij} dt_j\right) \tag{13.62}$$

from Eqn (13.58). Substituting in the equal-revenue constraint, Eqn (13.61), and recalling that $\partial L/\partial t_k = -\sum_{i=1}^{N} t_i M_{ik}$ yields:

$$dL = \left[\frac{\partial L}{\partial t_k}\left(-\frac{\frac{\partial T}{\partial t_j}}{\frac{\partial T}{\partial t_k}}\right) dt_j + \frac{\partial L}{\partial t_j} dt_j\right] \tag{13.63}$$

Rearranging terms:

$$\frac{dL}{dt_j} = \left[\frac{\partial L}{\partial t_k}\left(-\frac{\frac{\partial T}{\partial t_j}}{\frac{\partial T}{\partial t_k}}\right) + \frac{\partial L}{\partial t_j}\right] = \left[\frac{\partial L}{\partial t_k}\left(+\frac{dt_k}{dt_j}\bigg|_{R=\overline{R}}\right) + \frac{\partial L}{\partial t_j}\right] \tag{13.64}$$

Equation (13.64) gives an entirely plausible result. The change in loss from increasing one tax (say, t_j) and lowering another tax (say, t_k) to keep the total tax revenue constant is a linear combination of the marginal losses from changing t_k and t_j individually. The marginal loss for the revenue-compensating tax, t_k, is weighted by the amount that t_k must be changed per unit change in t_j in order to keep revenue unchanged. Put another way, the second term on the RHS of Eqn (13.64) measures the direct effect on loss because of a change in t_j. The first term measures the indirect effect on loss working through the required change in t_k in response to dt_j so that $dT = 0$. In the two-tax case, then, Eqn (13.64) gives an exact expression for the change in loss arising from a "small" equal-revenue tax substitution.

When more than two taxes change, an infinite number of combinations for dt can satisfy Eqn (13.59). The natural way to proceed in this case is to impose values on all but one of the tax changes, use Eqn (13.59) to solve for the remaining tax change, and then substitute for dt in Eqn (13.58).

Other than the obvious point that, given approximately equal-revenue effects, taxes that generate small changes in loss should replace taxes that generate large changes in loss to reduce loss, equations such as Eqn (13.64) are not particularly illuminating for policy purposes. Equation (13.64) can yield some interesting results, however, with additional restrictions added to the model in the form of limited possibilities for substitution and/or limitations in the number of taxed goods.

The Corlett and Hague Analysis

Corlett and Hague presented one of the more famous exercises along these lines, and one of the first. They examined the efficiency implications of moving from equal proportional taxes on two goods in the context of a three-good economy in which leisure is the third good and is incapable of being taxed.[37] Label the two goods k and j, and let good 1 be leisure, the untaxed numeraire ($q_1 \equiv 1 \equiv p_1$; $t_1 \equiv 0$). Assume initially that $t_j = \alpha \overline{p}_j$ and $t_k = \alpha \overline{p}_k$, with a the equal proportional rate of tax. Consider the efficiency implications of a marginal increase in t_j coupled with a marginal decrease in t_k that holds revenue constant—that is, an equal-revenue movement away from proportionality.

With proportional taxes a,

$$q_j = \overline{p}_j + \alpha \overline{p}_j = (1 + \alpha)\overline{p}_j \tag{13.65}$$

$$q_k = \overline{p}_k + \alpha \overline{p}_k = (1 + \alpha)\overline{p}_k \tag{13.66}$$

Substituting the expressions for marginal loss into Eqn (13.64) yields:

$$\frac{dL}{dt_j} = -\left[+\sum_{i=k,j} t_i M_{ik}\left(\frac{dt_k}{dt_j}\bigg|_{R=\overline{R}}\right) + \sum_{i=k,j} t_i M_{ij}\right] \tag{13.67}$$

To replace the terms $\sum_{i=k,j} t_i M_{ik}$ and $\sum_{i=k,j} t_i M_{ij}$, make use of the homogeneity condition and the symmetry of the Slutsky derivatives:

$$\sum_{i=1}^{3} q_i M_{ik} = \sum_{i=1}^{3} q_i M_{ij} = 0 \tag{13.68}$$

37. See Corlett and Hague (1953–1954). See also Diamond and McFadden (1974).

Rewrite $\sum_{i=1}^{3} q_i M_{ik} = 0$ as:

$$M_{1k} = -q_j M_{jk} - q_k M_{kk}, \quad \text{with } q_1 \equiv 1 \tag{13.69}$$

But, from Eqns (13.65) and (13.66)

$$M_{1k} = -\left[(1+\alpha)\overline{p}_j M_{jk} + (1+\alpha)\overline{p}_k M_{kk}\right] \tag{13.70}$$

Furthermore,

$$t_j M_{jk} + t_k M_{kk} = \alpha \overline{p}_j M_{jk} + \alpha \overline{p}_k M_{kk} \tag{13.71}$$

Hence, from Eqns (13.70) and (13.71),

$$\sum_{i=k,j} t_i M_{ik} = -\left(\frac{\alpha}{1+\alpha}\right) M_{1k} \tag{13.72}$$

Similarly,

$$\sum_{i=k,j} t_i M_{ij} = -\left(\frac{\alpha}{1+\alpha}\right) M_{1j} \tag{13.73}$$

Therefore, Eqn (13.67) becomes:

$$\frac{dL}{dt_j} = +\left(\frac{\alpha}{1+\alpha}\right)\left(M_{1k}\frac{dt_k}{dt_j}\bigg|_{R=\overline{R}} + M_{1j}\right) \tag{13.74}$$

Next, totally differentiate M_1, the demand for leisure, with respect to t_j, subject to the total revenue constraint and $q_1 \equiv 1$:

$$\frac{dM_1}{dt_j} = M_{1k}\frac{dt_k}{dt_j}\bigg|_{R=\overline{R}} + M_{1j} \tag{13.75}$$

Substituting Eqn (13.75) into (13.74) yields:

$$\frac{dL}{dt_j} = +\left(\frac{\alpha}{1+\alpha}\right)\frac{dM_1}{dt_j} \tag{13.76}$$

Thus, if leisure decreases (work increases) in response to the changes in $t_j(+)$ and $t_k(-)$, loss decreases, in which case equal proportional taxes are dominated by a system of nonproportional taxes on goods k and j.[38]

Whether or not dM_1/dt_j is negative depends on the Slutsky substitution terms M_{ij} and M_{ik}. To see this, write Eqn (13.75) as:

$$\frac{dM_1}{dt_j} = M_{1k}\left[-\frac{\frac{\partial T_j}{\partial t_j}}{\frac{\partial T_k}{\partial t_k}}\right] + M_{1j} \tag{13.77}$$

$$\frac{dM_1}{dt_j} = M_{1k}\left[-\frac{M_j + \sum_{i=k,j} t_i M_{ij}}{M_k + \sum_{i=k,j} t_i M_{ik}}\right] + M_{1j} \tag{13.78}$$

Substitute Eqns (13.72) and (13.73) into Eqn (13.78) to obtain:

$$\frac{dM_1}{dt_j} = M_{1k}\left[-\frac{\left(M_j - \left(\frac{\alpha}{1+\alpha}\right)M_{1j}\right)}{\left(M_k - \left(\frac{\alpha}{1+\alpha}\right)M_{1k}\right)}\right] + M_{1j} \tag{13.79}$$

Equation (13.79) assumes proportional taxes initially. Placing Eqn (13.79) over a common denominator and rearranging terms yields:

$$\frac{dM_1}{dt_j} = \frac{-M_j M_{1k} + M_k M_{1j}}{\left[M_k - \left(\frac{\alpha}{1+\alpha}\right)M_{1k}\right]} \tag{13.80}$$

Multiplying the first term in the numerator by M_k/M_k and the second term by M_j/M_j yields:

$$\frac{dM_1}{dt_j} = \frac{M_k M_j}{\left(M_k - \left(\frac{\alpha}{1+\alpha}\right)M_{1k}\right)} \cdot \left(\frac{M_{1j}}{M_j} - \frac{M_{1k}}{M_k}\right) \tag{13.81}$$

Assuming the first term on the RHS of Eqn (13.81) is positive, the sign of Eqn (13.81) depends upon the relative magnitudes of M_{1j}/M_j and M_{1k}/M_k. Consider the various possibilities. With $M_{11} < 0$, one possibility is $M_{1k} > 0$ and $M_{1j} < 0$. In the Slutsky sense, goods k and 1 are substitutes; goods j and 1, complements. If this is the case, then $dM_1/dt_j < 0$ as required for a decrease in loss. Hence, the government should raise the tax on the good that is complementary with leisure. If $M_{ik} < 0$ and $M_{ij} > 0$ then t_k should be raised.[39] If both are substitutes, such that M_{1k} and $M_{ij} > 0$, then Eqn (13.81) implies raising the tax on the good relatively more complementary (less substitutable) with leisure—for example, raising t_j if, roughly speaking, $M_{1j} < M_{1k}$, and vice versa. The only other possibility in a three-good world is for one of the goods to be a Slutsky substitute for leisure (say, $M_{1k} > 0$), while the other is neither a substitute nor a complement ($M_{1j} = 0$). In this case, the tax should be increased on the good for which the cross-price derivative is zero, since it is *relatively* more complementary with leisure (both goods cannot be Slutsky complements, since $\sum_{k=1}^{3} q_k M_{1k} = 0$ from homogeneity of the compensated demand functions and $M_{11} < 0$).

Note, finally, that the Corlett–Hague analysis applies, strictly speaking, only for small changes in taxes. Using the homogeneity conditions to obtain expressions in terms of M_{1k} and M_{1j} requires evaluating all demand relationships (M_k, M_{jk}, etc.) at the original proportional tax prices. The

38. By similar manipulations it can be demonstrated that $dL/dt_k = +\left(\frac{\alpha}{1+\alpha}\right)(dM_1/dt_k)$. Hence, if either dM_1/dt_k or dM_1/dt_j is negative, nonproportional taxes dominate proportional taxes.

39. If the analysis is carried out with respect to dt_k, the equation replacing Eqn (13.81) would be

$$\frac{dM_1}{dt_k} = \frac{M_k M_j}{\left[M_j - \left(\frac{\alpha}{1+\alpha}\right)M_{1j}\right]} \cdot \left(\frac{M_{1k}}{M_k} - \frac{M_{1j}}{M_j}\right) \tag{13.81n}$$

larger the tax changes, the more inaccurate this evaluation becomes. There are no longer any simple relationships between M_{1k} and $\sum_{i=k,j} t_i M_{ik}$ or between M_{1j} and $\sum_{i=k,j} t_i M_{ij}$.

REFERENCES

Atkinson, A., Stiglitz, J., April 1972. The structure of indirect taxation and economic efficiency. Journal of Public Economics Vol. 1 (1), 97–119.

Baumol, W., Bradford, D., June 1970. Optimal departures from marginal cost pricing. American Economic Review Vol. 60 (3), 265–283.

Besley, T., Jewitt, I., September 1995. Uniform taxation and consumer preferences. Journal of Public Economics Vol. 58 (1), 73–84.

Blumenthal, M., Slemrod, J., June 1992. The compliance cost of the U.S. individual income tax system: a second look after tax reform. National Tax Journal Vol. 45 (2), 185–202.

Blundell, R., MaCurdy, T., 1999. Labor supply: a review of alternative approaches. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3. Elsevier Science, Burlington (Chapter 27).

Blundell, R., Duncan, A., Meghir, C., July 1998. Estimating labor supply responses using tax reforms. Econometrica Vol. 66 (4), 827–861.

Bradford, D., Rosen, H., May 1976. The optimal taxation of commodities and income. American Economic Association Papers and Proceedings Vol. 66 (2), 94–101.

Browning, E., 1975. The excess burden of excise versus income taxes: a simplified comparison. Public Finance Vol. 30 (3), 445–451.

Browning, E., April 1976. The marginal cost of public funds. Journal of Political Economy Vol. 84 (2), 283–298.

Corlett, W., Hague, D., 1953–1954. Complementarity and the excess burden of taxation. Review of Economic Studies 21 (1), 21–30.

Diamond, P.A., Mirrlees, J., March, June 1971. Optimal taxation and public production (2 parts; Part 1: production efficiency, Part II: tax rules). American Economic Review Vol. 61 (1), 8–27; Vol. 61 (3:1), 261–278.

Diamond, P.A., McFadden, D., February 1974. Some uses of the expenditure function in public finance. Journal of Public Economics Vol. 3 (1), 3–21.

Dixit, A., February 1975. Welfare effects of tax and price changes. Journal of Public Economics Vol. 4 (2), 103–123.

Dixit, A., Munk, K., August 1977. Welfare effects of tax and price changes: a correction. Journal of Public Economics Vol. 8 (1), 103–107.

Feldstein, M., June 1995. The effects of marginal tax rates on taxable income: a panel study of the 1986 Tax Reform Act. Journal of Political Economy Vol. 103 (3), 551–572.

Feldstein, M., November 1999. Tax avoidance and the deadweight loss of the income tax. Review of Economics and Statistics Vol. 81 (4), 674–680.

Friedman, M., February 1952. The welfare effects of an income and an excise tax. Journal of Political Economy Vol. 60 (1), 25–33.

George, H., 1914. Progress and Poverty. Doubleday & Co., New York. Book VIII.

Green, J., Sheshinski, E., March 1979. Approximating the efficiency gains of tax reforms. Journal of Public Economics Vol. 11 (2), 179–195.

Gruber, J., Saez, E., April 2002. The elasticity of taxable income: evidence and implications. Journal of Public Economics Vol. 84 (1), 1–32.

Harberger, A., 1964a. Taxation, resource allocation and welfare. In: The Role of Direct and Indirect Taxes in the Federal Revenue System. The National Bureau of Economic Research and The Brookings Institution, Princeton University Press, Princeton, NJ.

Harberger, A., May 1964b. The measurement of waste. American Economic Association Papers and Proceedings Vol. 54 (3), 58–76.

Harberger, A., September 1971. Three basic postulates for applied welfare economics. Journal of Economic Literature Vol. 9 (3), 785–797.

Harberger, A., 1974. Taxation and Welfare. Little, Brown and Co., Boston, MA.

Hausman, J., 1981. Labor Supply. In: Aaron, H., Pechman, J. (Eds.), How Taxes Affect Economic Behavior. The Brookings Institution, Washington, DC.

Hausman, J., Rudd, P., May 1984. Family labor supply with taxes. The American Economic Association Papers and Proceedings Vol. 74 (2), 242–248.

Hotelling, H., July 1938. The general welfare in relation to problems of taxation and of railway and utility rates. Econometrica Vol. 6 (3), 242–269.

Kahn, A., 1970. The Economics of Regulation: Principles and Institutions, vol. 1. Wiley, New York, 144–145.

Kotlikoff, L., December 1984. Taxation and savings: a neoclassical perspective. Journal of Economic Literature Vol. 22 (4), 1576–1629.

MaCurdy, T., May 1992. Work disincentive effects of taxes: a reexamination of some evidence. American Economic Review Vol. 82 (2), 242–249.

Mirrlees, J., April 1971. An exploration in the theory of optimum income taxation. Review of Economic Studies Vol. 38 (2), 175–208.

Ramsey, F.P., March 1927. A contribution to the theory of taxation. Economic Journal Vol. 37 (145), 47–61.

Richter, D., October 1977. Games pythagoreans play. Public Finance Quarterly Vol. 5 (4), 495–515.

Summers, L., September 1981. Capital taxation and accumulation in a life cycle growth model. American Economic Review Vol. 71 (4), 533–544.

Samuelson, P.A., July 1986. Theory of optimal taxation. Journal of Public Economics Vol. 30 (2), 137–143.

Sandmo, S., July–August 1976. Optimal Taxation: An Introduction to the Literature. Journal of Public Economics Vol. 6 (1-2), 37–54.

Slemrod, J., May 1992. Do Taxes Matter? Lessons from the 1980s. American Economic Review Vol. 82 (2), 250–256.

Chapter 14

The Second-Best Theory of Taxation with General Production Technologies and Many Consumers

Chapter Outline

The second-best analysis of Chapter 13 must be extended in two directions to make it more responsive to real-world economies. One is to incorporate general production technologies, with increasing cost production-possibilities frontiers. Another is to consider the case of many consumers with different tastes and different marginal social welfare weights. Neither extension is analytically trivial.

With general technologies, producer prices vary as government policy variables move society along (or inside of) its production-possibilities frontier. Also, pure economic profits or losses are possible and have to be accounted for in a general equilibrium framework. As a consequence of these features, the marginal loss from taxation depends on production derivatives as well as consumption derivatives.

The many-consumer economy brings the social welfare function back into the analysis in a fundamental way, such that distinctions between the equity and efficiency implications of government policy become blurred. In addition, the concept of a general aggregate income measure of tax loss becomes problematic.

The modeling implications of either extension are sufficiently complex that Chapter 14 considers each separately before combining them into a fully general model. The first section of the chapter reworks two of the main results of Chapter 13 in the context of a one-consumer, general technology economy. The second section considers the many-consumer economy with fixed producer prices. The third and final section then presents the full general model and emphasizes how the results of Chapter 13 must be modified to accommodate a more realistic economic environment.

A ONE-CONSUMER ECONOMY WITH GENERAL TECHNOLOGY

Dead-Weight Loss from Taxation

Replacing the assumption of linear technology with the more realistic assumption of general technology affects the analysis of tax loss in two ways. The most direct implication is that production terms enter into the loss function in a nontrivial manner. In addition, general technology reintroduces market clearance explicitly into the analysis because the full set of market clearance equations is necessary to determine the relationship between producer and consumer prices. Each point deserves careful attention.

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00014-5

Pure Profits and Losses

With linear technology, the loss resulting from a vector of commodity taxes is defined as the lump-sum income necessary to keep the consumer indifferent to the taxes less the tax revenue collected at the compensated equilibrium and returned lump sum to the consumer, or $L(\vec{\mathbf{t}}) = M(\vec{\mathbf{q}}; \overline{U}) - \sum_{i=2}^{N} t_i X_i^C(\vec{\mathbf{q}}; \overline{U})$. There is no need to keep track of production because as society moves along a linear production frontier, there can never be any pure profits in the system that could also be given to the consumer. If, as we assumed, the competitively determined producer prices for the goods and factors generate no pure profits in the initial equilibrium, then there can never be pure profits because these prices never change. With general technologies, however, the competitively determined producer prices may well generate pure profits and losses, both at the initial zero-tax equilibrium and at the final with-tax equilibrium, and the pure profits may vary from one equilibrium to another. Consider, for example, the one-input, one-output, decreasing-returns-to-scale technology depicted in Fig. 14.1, in which input X_2 (measured negatively) produces output X_1.

The competitive price ratio P_{X_2}/P_{X_1} at the initial no-tax equilibrium A equals the slope of the line ab. Notice that at these prices the factor payments $P_{X_2} \cdot X_2$ do not exhaust the product $P_{X_1} \cdot X_1$. The firm earns pure profits equal to 0 a (in units of X_1) which presumably accrue to the single consumer. Note, also, that the value of the pure profits changes as society moves along the frontier in response to commodity taxes. As a result, loss must be reinterpreted more generally as the lump-sum income necessary to keep the consumer indifferent to the new consumer prices less all sources of lump-sum income available to the consumer at the new with-tax equilibrium. These include both the tax revenue that is returned lump sum and any pure profits existing at the new equilibrium.

Figure 14.2 shows the dead-weight loss from a unit tax on X_1 that changes the slope of the consumer's budget line to that of line segment cd. The income necessary to compensate the consumer for the new price vector is 0c (in units of X_1). The tax revenue collected and returned at the compensated equilibrium is cg, equal to the difference between the consumer prices (slope of cd) and the producer prices (slope of ef or gh) at the compensated equilibrium D, projected back to the X_1 axis. But loss is no longer the difference, 0g, because production at the compensated equilibrium gives rise to pure profits equal to 0e at the net-of-tax producer prices. These profits are also available to the consumer. Hence, the consumer's loss is only eg, equal to the difference between the consumer's required lump-sum compensation and the lump-sum income received from all sources within the economy. Notice that eg also equals the difference between the amount of X_1 required to compensate the consumer at D, less the amount of X_1 society is able to produce at E given the compensated supply of X_2, X_2^C.

The first requirement for retaining the loss-minimizing approach in general equilibrium analysis, then, is to develop a valid production relationship, specified in terms of producer prices, that measures the pure economic profits in the economy for any given vector of production prices. The proper analytical construct is the general equilibrium profit function. Assuming perfectly competitive goods and factor markets, the profit function $\pi(\vec{\mathbf{p}}) = \sum_{i=1}^{N} p_i Y_i(\vec{\mathbf{p}})$ is derived by assuming that a planner maximizes aggregate profits at fixed producer prices subject to the aggregate production-possibilities frontier $f(\vec{\mathbf{Y}}) = 0$. The resulting aggregate goods supply and input demand functions $Y_i(\vec{\mathbf{p}})$ are then substituted back into the profit function $\sum_{i=1}^{N} p_i Y_i$. Analogous to the consumer's expenditure function, $\partial\pi(p)/\partial p_k = Y_k(\vec{\mathbf{p}})$, the

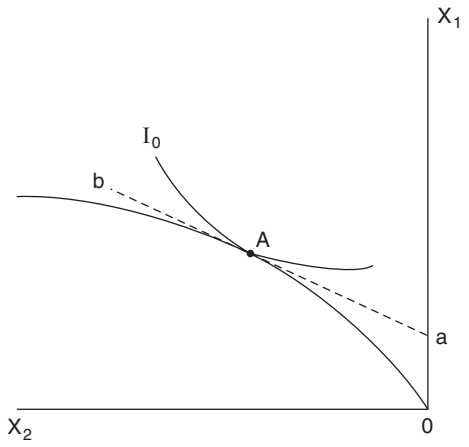

FIGURE 14.1

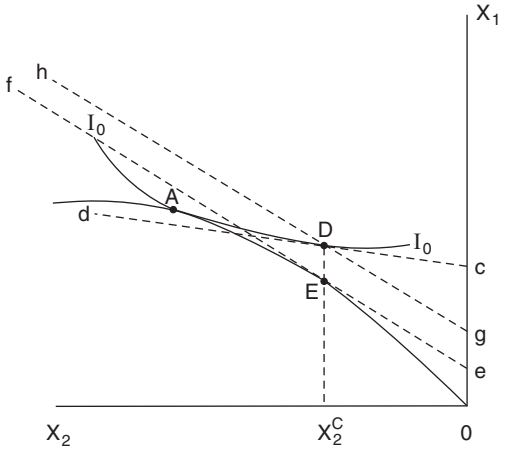

FIGURE 14.2

supply of (demand for) the kth good (factor), a property known as Shepard's lemma.[1]

The general equilibrium profit function $\pi(\vec{\mathbf{p}})$ incorporates all relevant aspects of production for the economy. Therefore, the expression

$$L(\vec{\mathbf{t}}) = M(\vec{\mathbf{q}}; \overline{U}) - \sum_{i=1}^{N} t_i M_i(\vec{\mathbf{q}}; \overline{U}) - \pi(\vec{\mathbf{p}}) \quad (14.1)$$

is a valid general equilibrium expression for the deadweight loss resulting from any given vector of taxes, $\vec{\mathbf{t}}$, assuming competitive market structures, general production technology, and tax revenues measured at the new with-tax compensated equilibrium. The expenditure function $M(\vec{\mathbf{q}}; \overline{\mathrm{U}})$ measures the income necessary to compensate the consumer, and the final two expressions measure the (lump-sum) income actually available at the new compensated equilibrium.[2] As before, $\vec{\mathbf{q}}$ is the vector of gross-of-tax consumer prices, $\vec{\mathbf{p}}$ is the vector of net-of-tax producer prices, and $\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$.

Market Clearance

Equation (14.1) is not a complete general equilibrium specification of the economy, however, unlike the loss expression with a linear technology. Although $M(\vec{\mathbf{q}}; \overline{U})$ completely specifies the preferences of the consumer and $\pi(\vec{\mathbf{p}})$ completely specifies the production technology under perfect competition, $L(t)$ does not incorporate market clearance. Recall that market clearance was implicit with linear technology. There was only one consumer, and aggregate production and producer prices were fixed. Therefore, once $\vec{\mathbf{t}}$ was specified, $\vec{\mathbf{q}}$ was determined through the identities $\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$. There is still only one consumer and aggregate production with general technology, but the crucial difference is that producer prices are no longer fixed. In general, goods supply curves (input demand curves) are upward (downward) sloping so that any given producer price $\mathrm{p_i}$ is now a function of the entire vector of taxes, $\vec{\mathbf{t}}$, that is, $p_i = p_i(\vec{\mathbf{t}})$. To see this, consider the response to a tax, t_k, in both the market for k and the market for some other good, i. In the market for good k, depicted in Fig. 14.3, the tax t_k generates a new equilibrium X_k^T, with new consumer and producer prices q_k^1 and p_k^1. Similarly, in some other market i, if D_i shifts in response to the tax t_k (as pictured in Fig. 14.4, the goods i and k are substitutes), both the consumer and the

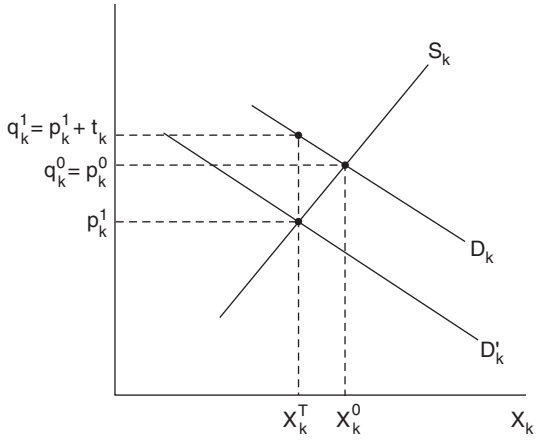

FIGURE 14.3

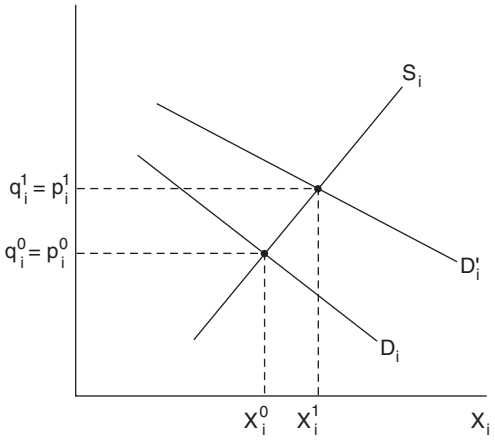

FIGURE 14.4

producer prices change from $q_i^0 = p_i^0$ to $q_i^1 = p_i^1$.[3] With linear technologies, in contrast, all supply (input demand) curves are horizontal so that a tax t_k could only change the consumer price q_k by the full amount of the tax. p_k could not change, nor could any other producer or consumer price, even if the tax t_k caused demand shifts in these other markets.

With general technology, therefore, market clearance relationships of the form

$$M_i(\vec{\mathbf{q}}; \overline{U}) = M_i[\vec{\mathbf{p}}(\vec{\mathbf{t}}) + \vec{\mathbf{t}}; \overline{U}] = Y_i[\vec{\mathbf{p}}(\vec{\mathbf{t}})] \quad i = 1, \ldots, N \quad (14.2)$$

become necessary to determine the vector of producer prices $\vec{\mathbf{p}}$ for any given vector of taxes $\vec{\mathbf{t}}$. Once $\vec{\mathbf{p}}$ has

1. Shepard's lemma is derived as follows: $\partial\pi(p)/\partial p_k = \partial \sum_i p_i Y_i(p)/\partial p_k = Y_k(p) + \sum_i p_i \partial Y_i/\partial p_k$. But $p_i = \lambda f_i$ from the first-order conditions for profit maximization. Therefore, $\partial\pi(p)/\partial p_k = Y_k + \lambda \sum_i f_i \partial Y_i/\partial p_k$. The second expression equals zero with $f(Y) = 0$, so that $\partial\pi/\partial p_k = Y_k$.

2. As in Chapter 13, $\overline{U}$ is arbitrarily set at U^0, the utility level attainable without any taxation. Refer to Chapter 13 for a discussion of this choice.

3. This diagrammatic analysis ignores further price changes as D_k shifts in response to the changes in q_i, which changes q_k and further shifts D_i, and so on.

been determined through these relationships, $\vec{\mathbf{q}}$ is determined by the identities $\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$.

The market clearance relationships, Eqn (14.2), can be incorporated into the analysis in one of two ways. One possibility is to replace either M_i or Y_i in Eqn (14.1), the expression for loss. The other choice is to keep Eqn (14.1) exactly as it is and use the market clearance relationships to simplify derivatives of $L(t)$. We will use the second method throughout the chapter.

There is one additional complication. The loss expression, Eqn (14.1), is specified in terms of compensated goods demands and factor supplies. To be consistent, the market clearance relationships, Eqn (14.2), must also be specified in terms of the compensated demands and supplies, the $M_i(\vec{\mathbf{q}}; \overline{U})$, as written. But, market clearance cannot possibly hold in terms of all N compensated supplies and demands, because the consumer would not suffer any loss as a result of the commodity taxes if society could provide the full vector of compensated supplies and demands. As noted above, the compensated equilibrium for the consumer in Fig. 14.2, point D (the $M_i(\vec{\mathbf{q}}; \overline{U})$), is not attainable with the given production technology. Production at the compensated equilibrium (the $Y_i(\vec{\mathbf{p}})$) is represented by point E. Thus, specifying that $M_i = Y_i$, for all $i = 1, \ldots, N$, would require that E and D coincide, which cannot possibly occur if there is dead-weight loss.

A natural resolution of the market clearance problem is to impose market clearance on all but one of the goods and factors, say, the first, and assume that compensation occurs through this good. It is also natural to let good one serve as the untaxed numeraire, with $q_1 \equiv p_1 \equiv 1$, and $t_1 = 0$. From the discussion of loss in Chapter 13, we know that loss requires a change in relative prices, and setting $t_1 = 0$ ensures that any tax vector must change the vector of relative prices. Moreover, with $q_1 \equiv p_1 \equiv 1$, units of X_1 can be interpreted as units of purchasing power. These assumptions are consistent with Fig. 14.2 in which loss is depicted as eg (or DE), equal to the units of X_1 demanded at the compensated equilibrium less the amount of X_1 actually produced given the producer prices at that equilibrium. In fact, given that $M_i = Y_i$, $i = 2, \ldots, N$, and $q_1 \equiv p_1 \equiv 1$, $t_i \equiv 0$, the loss from taxation can be written as

$$L(\vec{\mathbf{t}}) = M_1(\vec{\mathbf{q}}; \overline{U}) - Y_1(\vec{\mathbf{p}}) \tag{14.3}$$

the amount of excess demand for good 1 at the compensated equilibrium. Equations (14.3) and (14.1) are equivalent specifications of dead-weight loss for analytical purposes.

The choice of the untaxed numeraire and the corresponding uncleared market is immaterial since it has no effect on the vector of relative prices and therefore on the compensated equilibrium. The numerical value of loss changes as it gets expressed in different units of the chosen numeraire good, but not the compensated equilibrium. Nonetheless, the discussion so far suggests that the concept of dead-weight loss in general technology is not as useful as we might like. The problem is that the vector of (relative) prices $\vec{\mathbf{p}}$ solved through the $(N - 1)$ market clearing equations differs, in general, from the actual $\vec{\mathbf{p}}$ observed in the economy in response to any given $\vec{\mathbf{t}}$.

The issue of compensation thus points to a dilemma between actual and compensated equilibria. If we want to define dead-weight loss for a general technology economy, all components of the loss function must be defined in terms of the compensated equilibrium resulting from any given tax vector. If the actual and compensated equilibria are mixed together by, say, returning the actual tax revenue collected, then the loss minimization specification of a particular problem does not generate the same analytical results as a welfare maximum specification. Thus, to be an entirely consistent general equilibrium exercise, the conceptual loss experiment must assume, in effect, that the consumer is actually compensated by some outside agent and that the compensation takes place in terms of some particular good. Were such compensation to occur, the price vector $\vec{\mathbf{p}}$ solved for by this experiment would be the actual price vector observed in the economy. The dilemma arises because the compensation does not actually occur, so that the price vector $\vec{\mathbf{p}}$ resulting from the conceptual loss experiment is neither the same as, nor does it bear any necessary relationship to, the actual $\vec{\mathbf{p}}$ resulting from any given pattern of taxes, $\vec{\mathbf{t}}$. The actual $\vec{\mathbf{p}}$ are irrelevant to a carefully designed conceptual experiment defining dead-weight loss.

There appears to be no way out of this dilemma as long as one remains interested in defining a legitimate loss measure. One can avoid the dilemma by modeling all second-best tax issues as welfare maximization problems using the indirect utility function, in which case everything is defined in terms of the actual post- and pretax equilibria. There is no need to define a loss function to analyze second-best tax (or expenditure) issues. The loss minimization framework is compelling, however, since dead-weight loss has been the traditional notion of tax inefficiency.

This dilemma does not arise with the same force under linear technologies because producer prices are fixed. The conceptual loss experiment still involves the compensated rather than actual tax collections and is therefore somewhat removed from the actual equilibrium. But, the loss experiment at least employs the observed vector of prices, both $\vec{\mathbf{q}}$ and $\vec{\mathbf{p}}$, for any given vector of tax rates, $\vec{\mathbf{t}}$. The fixed vector $\vec{\mathbf{p}}$ is the same in each equilibrium.

Marginal Loss: General Technology

With these comments in mind, we can analyze the loss from taxation in a one-consumer economy with general technology. Begin by computing the marginal loss with

respect to the kth tax, t_k. Use the loss expression, Eqn (14.1), $L(\vec{\mathbf{t}}) = M(\vec{\mathbf{q}}; \overline{U}) - \sum_{i=1}^{N} t_i M_i(\vec{\mathbf{q}}; \overline{U}) - \pi(\vec{\mathbf{p}})$, along with the $(N-1)$ market clearance relationships:

$$M_i(\vec{\mathbf{q}}; \overline{U}) = Y_i(\vec{\mathbf{p}}) \quad i = 2, \ldots, N \tag{14.4}$$

and the pricing identities

$$q_i = p_i + t_i \quad i = 2, \ldots, N \tag{14.5}$$

plus

$$q_1 \equiv p_1 \equiv 1, \quad t_1 \equiv 0 \tag{14.6}$$

Given that $p_i = p_i(t), \quad i = 2, \ldots, N, \quad \partial\pi/\partial p_i \equiv Y_i$, and noting that $p_1 \equiv 1$,[4]

$$\frac{\partial L}{\partial t_k} = M_k + \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} - M_k - \sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) - \sum_{i=2}^{N} Y_i \frac{\partial p_i}{\partial t_k} \tag{14.7}$$

$$\frac{\partial L}{\partial t_k} = \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} - \sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) - \sum_{i=2}^{N} Y_i \frac{\partial p_i}{\partial t_k} \tag{14.8}$$

Equation (14.8) can be simplified further by means of the market clearance equations, Eqn (14.4). Multiplying Eqn (14.4) by $\partial p_i/\partial t_k$ yields

$$M_i \frac{\partial p_i}{\partial t_k} = Y_i \frac{\partial p_i}{\partial t_k} \quad i = 2, \ldots, N \tag{14.9}$$

Next, sum Eqn (14.9) over all $(N-1)$ relationships to obtain

$$\sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} = \sum_{i=2}^{N} Y_i \frac{\partial p_i}{\partial t_k} \tag{14.10}$$

Hence, Eqn (14.8) simplifies to

$$\frac{\partial L}{\partial t_k} = -\sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) \tag{14.11}$$

Equation (14.11) is similar to Eqn (13.6), the expression for marginal loss with linear technology. With $q_i = p_i + t_i$, for $i = 2, \ldots, N$, Eqn (14.11) can be rewritten as

$$\frac{\partial L}{\partial t_k} = -\sum_{i=2}^{N} t_i \frac{\partial X_i^{\text{comp}}}{\partial q_j} \frac{\partial q_j}{\partial t_k} = -\sum_{i=2}^{N} t_i \frac{\partial X_i^{\text{comp}}}{\partial t_k} \quad k = 2, \ldots, N \tag{14.12}$$

Once again, we see that marginal loss depends upon the pattern of existing taxes and the change in compensated demands (factors supplies) in response to the tax. With a linear technology, $\partial q_k = \partial t_k$ and $\partial q_j/\partial t_k = 0, j \neq k$, so that Eqn (13.6) is just a special case of the general expression (14.11). The major qualitative difference between the two expressions is that the derivative $\partial X_i^{\text{comp}}/\partial t_k$ depends on both consumption and production responses, since $\partial q_j/\partial t_k$ depends upon all the consumption and production elasticities through the $(N-1)$ market clearance equations, Eqn (14.4). This is hardly a trivial difference, of course.

To see the roles of the consumption and production derivatives, rewrite the $(N-1)$ market clearance relationships, Eqn (14.4), and the expression for loss, Eqn (14.11), in vector notation[5]:

$$dL = -(t')(M_{ij}) \left(\frac{dq}{dt} \right) (dt) \tag{14.13}$$

$$M_i(q; \overline{U}) = \pi_i(q - t) \tag{14.14}$$

where

(t) = the $(N-1) \times 1$ column vector

$$\begin{bmatrix} t_2 \\ \vdots \\ t_N \end{bmatrix}$$

(M_{ij}) = the $(N-1) \times (N-1)$ matrix:

$$\begin{bmatrix} M_{22} & \cdots & M_{2N} \\ \vdots & & \vdots \\ M_{N2} & \cdots & M_{NN} \end{bmatrix}$$

$\left(\frac{dq}{dt}\right)$ = the $(N-1) \times (N-1)$ matrix of differentials:

$$\begin{bmatrix} \frac{dq_2}{dt_2} & \cdots & \frac{dq_2}{dt_N} \\ \vdots & & \vdots \\ \frac{dq_N}{dt_2} & \cdots & \frac{dq_N}{dt_N} \end{bmatrix}$$

(dt) = the $(N-1) \times 1$ column vector of differentials:

$$\begin{bmatrix} dt_2 \\ \vdots \\ dt_N \end{bmatrix}$$

M_i = the $(N-1) \times 1$ column vector of compensated demands (factor supplies):

$$\begin{bmatrix} M_2 \\ \vdots \\ M_N \end{bmatrix}$$

4. $q_i = p_i(t) + t_i$; therefore, $\partial q_i/\partial t_k = \partial p_k/\partial t_k$, $i \neq k$. $\partial q_k/\partial t_k = \partial p_i/\partial t_k + \partial t_k/\partial t_k = \partial p_k/\partial t_k + 1$.

5. This technique was first demonstrated to us by Peter Diamond in a set of unpublished class notes.

π_i = the $(N-1) \times 1$ column vector of supplies (input demands):

$$\begin{bmatrix} Y_2 \\ \vdots \\ Y_N \end{bmatrix}$$

q, p = the $(N-1) \times 1$ column vectors of prices.

Totally differentiating Eqn (14.14) yields

$$M_{ij}\mathrm{d}q = Y_{ij}(\mathrm{d}q - \mathrm{d}t) \tag{14.15}$$

Solving for $\mathrm{d}q/\mathrm{d}t$ and substituting the notation X for M_i and Y for π_i yields

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \frac{-\left(\frac{\partial Y}{\partial p}\right)}{\left(\frac{\partial X}{\partial q} - \frac{\partial Y}{\partial p}\right)} \tag{14.16}$$

Substituting Eqn (14.16) into (14.13), the expression for loss becomes

$$\mathrm{d}L = -(t')(M_{ij})\left[\frac{-\frac{\partial Y}{\partial p}}{\frac{\partial X}{\partial q} - \frac{\partial Y}{\partial p}}\right]\mathrm{d}t \tag{14.17}$$

Hence, marginal loss depends upon both consumption and production derivatives.

As one additional comparison of marginal losses in general versus linear technology economies, consider the simple (and unlikely) case in which t_k is the only existing tax, only the kth prices, q_k and p_k, vary in response to a marginal change in the kth tax, t_k, and that all cross-price derivatives are zero.[6] With these assumptions, Eqn (14.13) (or Eqn (14.17)) simplifies to

$$\mathrm{d}L = -t_k \frac{\partial X_k}{\partial q_k}\frac{\mathrm{d}q_k}{\mathrm{d}t_k}\mathrm{d}tk = -t_k \frac{\partial X_k}{\partial q_k}\mathrm{d}q_k = -t_k \Delta X_k \tag{14.18}$$

The marginal loss occurs entirely within the market for good (factor) k and is approximately equal to the shaded trapezoidal area in Fig. 14.5. This area can be thought of as the combined (marginal) decrease in consumer's and producer's surplus from consuming and producing good k, where the former is measured with reference to the compensated demand for good k and the latter with reference to the generalized supply function $Y_k = \partial\pi(\vec{\mathbf{p}})/\partial p_k$. By contrast, with linear technology loss was approximated by a set of triangles, equal in each market to the loss in consumer's surplus measured with reference to the set of compensated demand (factor supply) curves. A generalized producer's surplus cannot arise in a linear technology with its perfectly elastic output supplies (input demands) at constant producer prices.

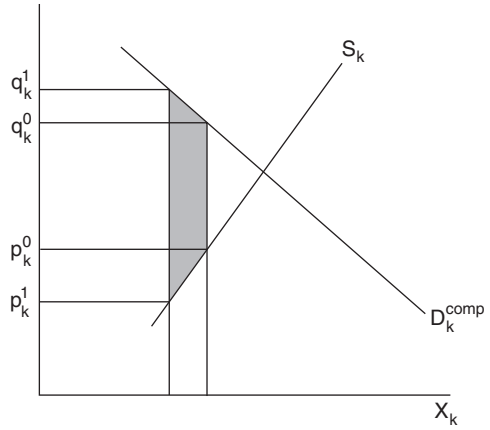

FIGURE 14.5

Optimal Commodity Taxation

One of the more important results in the allocational theory of taxation is that the equations describing the optimal pattern of commodity taxes in a one-consumer, general technology economy are identical to their linear technology counterparts if the technology exhibits constant returns to scale (CRS). The first-order conditions for optimal taxation continue to depend only upon compensated demand (factor supply) derivatives even though the marginal tax loss in general technology with CRS depends upon consumption and production derivatives. Having already discussed the notion of loss from taxation with general technologies, this result is easily derived.

The optimal commodity tax problem in a one-consumer general technology economy can be represented as

$$\min_{(t_k)} L(t) = M(\vec{\mathbf{q}}; \overline{\mathrm{U}}) - \sum_{i=2}^{N} t_i M_i - \pi(\vec{\mathbf{p}})$$

$$\text{s.t.} \sum_{i=2}^{N} t_i M_i(\vec{\mathbf{q}}; \overline{U}) = \overline{T}$$

along with the market clearance equations, Eqn (14.4), and the pricing identities, Eqns (14.5) and (14.6). Notice once again that tax revenue is measured at the compensated equilibrium. The corresponding Lagrangian is

$$\min_{(t_k)} \xi(t) = M(\vec{\mathbf{q}}; \overline{U}) - \sum_{i=2}^{N} t_i M_i - \pi(\vec{\mathbf{p}}) - \lambda\left(\sum_{i=2}^{N} t_i M_i(\vec{\mathbf{q}}; \overline{U}) - \overline{T}\right)$$

6. Assuming $M_{ij} = 0$, for all $i \neq j$ is improper, but the example is meant to be illustrative only. This analysis, along with the result, Eqn (14.17), can be found in Boadway (1975). Boadway derives Eqn (14.17) in a utility-maximizing framework.

The first order conditions are

$$\frac{\partial L}{\partial t_k} - \lambda \frac{\partial T}{\partial t_k} = 0 \quad k = 2, \ldots, N \tag{14.19}$$

and

$$\sum_{i=2}^{N} t_i M_i = T \tag{14.20}$$

But,

$$\frac{\partial L}{\partial t_k} = -\sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) \tag{14.21}$$

and

$$\frac{\partial T}{\partial t_k} = M_k + \sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) \tag{14.22}$$

Therefore, Eqn (14.19) becomes

$$(1+\lambda)\frac{\partial L}{\partial t_k} - \lambda M_k = 0 \quad k = 2, \ldots, N \tag{14.23}$$

or

$$-(1+\lambda)\left[\sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) \right] - \lambda M_k = 0 \quad k = 2, \ldots, N \tag{14.24}$$

Without imposing the assumption of CRS, all we can do is rewrite Eqn (14.24) in a form corresponding, but not identical, to the optimal tax rules for a linear technology:

$$\left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) = \sum_j \frac{\partial X_i^{\text{comp}}}{\partial q_j} \cdot \frac{\partial q_j}{\partial t_k} = \frac{\partial X_i^{\text{comp}}}{\partial t_k} \quad \text{all } i = 2, \ldots, N \tag{14.25}$$

Therefore, the first-order conditions, Eqn (14.24), are equivalent to

$$\frac{-\sum_{i=2}^{N} t_i \frac{\partial X_i^{\text{comp}}}{\partial t_k}}{M_k} = \frac{\lambda}{1+\lambda} = C \quad k = 2, \ldots, N \tag{14.26}$$

As was true with the expression for marginal loss, the linear technology rules, Eqn (13.30), are a special case of the general Eqn (14.26), with $\partial q_k = \partial t_k$ and $\partial q_j / \partial t_k = 0$, $j \neq k$. But, with general technology, the derivatives $\partial X_i^{\text{comp}} / \partial t_k$ in Eqn (14.26) refer to the general equilibrium changes in X_i^{comp} in response to the tax, which in turn depend upon the changes in the full set of producer and consumer prices as t_k changes. And, as demonstrated above, these price changes are functions of both demand and supply price derivatives through the market clearance equations.

It is not immediately obvious why the assumption of CRS should reduce the general Eqn (14.26) to their linear technology counterparts. After all, even with CRS the output supply (input demand) curves are generally upward (downward) sloping, which means that producer prices vary in response to variations in government taxes.[7] But the key is that there can be no pure profits in the economy with CRS technology and perfectly competitive market structures. $\pi(\vec{\mathbf{p}}) \equiv 0$, so that

$$\frac{\partial \pi}{\partial t_k} = \sum_{i=1}^{N} \pi_i \frac{\partial p_i}{\partial t_k} = 0$$

With $p_1 \equiv 1$ as the numeraire, $\sum_{i=1}^{N} \pi_i (\partial p_i / \partial t_k) = 0$ as well. But this implies, from market clearance, that

$$\sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} = \sum_{i=2}^{N} \pi_i \frac{\partial p_i}{\partial t_k} \equiv \sum_{i=2}^{N} Y_i \frac{\partial p_i}{\partial t_k} = 0 \tag{14.27}$$

Using this result, subtract $\lambda \sum_{i=2}^{N} M_i (\partial p_i / \partial t_k) (= 0)$ from Eqn (14.24), obtaining

$$\begin{aligned} &-(1+\lambda)\sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) \\ &-\lambda \left(M_k + \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} \right) = 0 \end{aligned} \tag{14.28}$$

Writing all $(N-1)$ of these equations in matrix notation,

$$-(1+\lambda)(t')(M_{ij})\left[1 + \left(\frac{\partial p}{\partial t} \right) \right] - \lambda (M_i') \left[I + \left(\frac{\partial p}{\partial t} \right) \right] = 0' \tag{14.29}$$

where

t, M_{ij}, and M_i are defined as above.
λ and $(1+\lambda)$ are scalars.
I is the $(N-1) \times (N-1)$ identity matrix.
$\left(\frac{\partial p}{\partial t} \right)$ = the $(N-1) \times (N-1)$ matrix of price derivatives:

$$\begin{bmatrix} \frac{\partial p_2}{\partial t_2} & \cdots & \frac{\partial q_2}{\partial t_N} \\ \vdots & & \vdots \\ \frac{\partial p_N}{\partial t_2} & \cdots & \frac{\partial p_N}{\partial t_N} \end{bmatrix}$$

0 = an $(N-1)$ column vector of zeros.

7. CRS generates an increasing-cost production-possibilities frontier so long as production of the various goods is unequally factor intensive, meaning that the optimal factor proportions across goods differ at the same relative factor prices.

Since $[I + (\partial p/\partial t)]$ is nonsingular, Eqn (14.29) implies

$$-(1+\lambda)(t')(M_{ij}) - \lambda(M_i') = 0' \tag{14.30}$$

Equation (14.30) holds for each of the $(N - 1)$ relationships. Hence,

$$-(1+\lambda)\sum_{i=2}^{N} t_i M_{ik} - \lambda M_k = 0 \quad k = 2, \ldots, N \tag{14.31}$$

Rearranging terms,

$$\frac{\sum_{i=2}^{N} t_i M_{ik}}{M_k} = -\frac{\lambda}{1+\lambda} = C \quad k = 2, \ldots, N \tag{14.32}$$

with C independent of k. Equation (14.32) is identical to Eqn (13.30). They imply that the pattern of optimal taxes depends only upon the compensated demand (factor supply) derivatives $\partial X_i^{\text{comp}}/\partial q_k$. Moreover, the equal percentage change interpretation applies to Eqn (14.32) exactly as it applies to Eqn (13.30).[8]

The assumption of CRS, then, greatly simplifies the application of second-best results. Admittedly, Eqn (14.32) would be difficult to apply in practice given our limited econometric knowledge of the relevant Slutsky substitution terms, but at least the general equilibrium supply responses to the tax can be ignored.

That the assumption of CRS for private production simplifies the optimal commodity tax rules is not unique to that problem. CRS tends to simplify all second-best results in both tax and expenditure theory. Whether CRS is an appropriate assumption for the US economy is an open question. CRS is often assumed for aggregate production in empirical analysis, but this is mostly because aggregate production data are collected by the government and are constructed under the assumption of CRS (exhaustion of the product). The same is true of production analysis at the two-digit industry level.

Another difficulty in applying this, and other, second best results is the market assumption of perfect competition. It is certainly violated for a number of important goods and services. At the same time, perfect competition is the natural default assumption for second-best public sector analysis. One could hardly attempt to model the various forms that market imperfections take industry by industry throughout the economy.

MANY-PERSON ECONOMIES: FIXED PRODUCER PRICES

Social Welfare Maximization versus Loss Minimization

Before 1970, public sector economists chose to analyze second-best tax theory almost exclusively within the context of one-consumer economies to highlight the efficiency aspects of that theory. The results derived in Chapters 13 and 14 provide a representative sampling of the received theory in the professional journals up to 1970. During the 1970s, however, a number of the leading public sector theorists—Boadway, Diamond, Feldstein, Green, Hartwick, and Mirrlees, to name just a few (Boadway, 1976; Diamond, 1975; Feldstein, 1972; Green, 1975; Hartwick, 1978; Mirrlees, 1975)—reworked second-best tax theory within the more realistic context of many-person economies. These first papers showed that it might not be very useful to consider efficiency aspects of various taxes independently of their distributional effects, at least not for the purposes of practical application. Economists had long known that distributional considerations would modify the standard one-consumer results of second-best tax theory, but the many-person models made it painfully obvious just how hopelessly intertwined distributional and efficiency terms become in many second-best tax (and expenditure) decision rules. This is especially disturbing because arbitrary assignment of the distributional weights embodied in an underlying social welfare function can generate quite different policy implications from these decision rules.

Along these same lines, it may not be very useful to think of the effects of distorting taxes in terms of deadweight loss, even though public sector economists have characterized distortion as loss since the very beginnings of the discipline. Unambiguous notions of efficiency loss involve the use of the expenditure function, which is best suited to one-consumer economies. Loss minimization and welfare minimization generate identical results in second-best analysis if the objective function is the welfare or loss of a single individual. In a many-person economy, however, loss minimization and social welfare maximization are no longer equivalent except under the highly restrictive assumptions that render the many-person economy essentially equivalent to the one-person economy. Indeed the concept of loss is not generally well defined in a many-person economy. This point is worth considering before analyzing a specific second-best problem in a many-person context.[9]

8. The independence of the optimal tax rules to supply responses with constant returns-to-scale production was first pointed out to us by Paul Samuelson in a set of unpublished class notes. Refer to Stiglitz and Dasgupta (1971), for an alternative derivation of this result.

9. This point was discussed in Chapter 4. Here we assume a CRS, zero-profit economy.

Loss measures using the expenditure function model the economy in terms of market prices. Therefore, loss can be directly compared with social welfare expressed in terms of each consumer's indirect utility function, $V^h(\vec{\mathbf{q}}; I^h)$. The indirect utility function is obtained by solving for the consumer's demand (input supply) functions $X_{hk} = X_{hk}(\vec{\mathbf{q}}; I^h)$ from utility maximization and substituting them for the arguments of the direct utility function $U^h(X_{hk})$. Let

$$W^*\left[V^h\left(\vec{\mathbf{q}}; I^h\right)\right] = V\left(\vec{\mathbf{q}}; I^1, \ldots, I^h, \ldots, I^H\right) \qquad (14.33)$$

represent the Bergson–Samuelson individualistic social welfare function expressed as a function of the vector of consumer prices, $\vec{\mathbf{q}}$, and the distribution of lump-sum incomes (I^1,…, I^H).

The problem is that, in general, there exists no aggregate expenditure function of the form $M(\vec{\mathbf{q}}; \overline{V}^1, \ldots, \overline{V}^H)$ corresponding to the social welfare function, which can be incorporated into a many-person loss measure, because there is no unambiguous method for specifying the vector of constant utilities, $\overline{V}_1, \ldots, \overline{V}_H$, to be inserted into M. Suppose, for example, that the government were to change the vector of consumer prices, $\vec{\mathbf{q}}$, by instituting a set of distorting taxes, $\vec{\mathbf{t}}$. A natural way of defining M would be to hold each consumer at his pretax utility level and ask how much lump-sum income in the aggregate would be required to do this given the new gross-of-tax consumer prices. In effect, each consumer would be fully compensated for the tax, with Eqn (14.33) evaluated at the pretax utility levels $(\overline{V}_0^1, \ldots, \overline{V}_0^H)$. Imagine that the government actually borrowed (at no cost) the required income from some third country and compensated each consumer. Clearly, this amount of income would differ from the income required to keep social welfare constant in response to the tax, because by returning each consumer to his pretax utility the government has foregone the possibility of exploiting differences in the social welfare weights $\partial W^*/\partial V^h$. By judiciously offering more income to people with high marginal social utilities and less to those with low marginal social utilities, the government can restore the pretax level of social welfare without necessarily returning each consumer to his pretax utility level. The only appropriate vector of utilities $(\overline{V}_1, \ldots, \overline{V}_H)$ to plug into M, therefore, is the vector of individual utilities that would exist once social welfare has been "compensated" at its pretax level, but there is no general method of solving for this vector. In particular, a many-person loss measure of the form

$$L(\vec{\mathbf{t}}) = \sum_{h=1}^{H} L^h(\vec{\mathbf{t}}) = \sum_{h=1}^{H}\left[M^h\left(\vec{\mathbf{q}}; \overline{U}^h\right) - \sum_{i=2}^{N} t_i X_{hi}\right],$$

the straight sum of each individual's loss, bears no necessary relationship to the social welfare function $V(\vec{\mathbf{q}}; I^1, \ldots, I^H)$.

One might think that weighting each $L^h(\vec{\mathbf{t}})$ by the marginal social welfare terms $\partial W^*/\partial V^h$ and defining aggregate loss as

$$L(\vec{\mathbf{t}}) = \sum_{h=1}^{H} \frac{\partial W^*}{\partial V^h} \cdot L^h(\vec{\mathbf{t}})$$

would be equivalent to Eqn (14.33), but that is not so. It turns out that the proper weighting scheme for individual losses is problem specific. Terms from second-best constraints must be incorporated into the vector of weights to make loss minimization equivalent to social welfare maximization.

The aggregate expenditure function is unambiguously defined for a many-consumer economy only if the economy is equivalent to a single-consumer economy in the sense that the level of social welfare is independent of any distributional considerations, including both the distribution of lump-sum income and the pattern of consumption (and factor supply) among the various consumers. Three sufficient conditions have been described that will generate one-consumer equivalence, two by Samuelson and one by Green, as follows (Samuelson, 1956; Green, 1975):

1. Lump-sum income is continuously and optimally redistributed in accordance with the interpersonal equity conditions of first-best social welfare maximization. That is, the social marginal utility of income is always equal for all consumers.
2. Consumers have identical and homothetic tastes so that for any given consumer price vector, $\vec{\mathbf{q}}$, and all lump-sum income distributions I^1, …, I^H, the aggregate Engel's (income–consumption) curves are straight parallel lines.
3. The covariance of person h's social marginal utility of income and his proportion of aggregate consumption of any one good (X_{hk}/X_k) is identical for all goods (and factors) $k = 1, \ldots, N$ (Green's condition).

Under any of these conditions, the function $V(\vec{\mathbf{q}}; I^1, \ldots, I^H)$ is equivalent to $V(\vec{\mathbf{q}}; I)$, which in turn is identical to the specification of indirect utility for a single consumer. Moreover, if social welfare can be expressed as $V(\vec{\mathbf{q}}; I)$, then the problem

$$\max_{(q)} V(\vec{\mathbf{q}}; I)$$

$$\text{s.t. } \vec{\mathbf{q}} \cdot \vec{\mathbf{X}} = 1$$

where

$\vec{\mathbf{X}}$ = the vector of aggregate quantities
I = aggregate lump-sum income

has the dual form

$$\max_{(X)} \vec{\mathbf{q}} \cdot \vec{X}$$

$$\text{s.t. } V = \overline{V}$$

The dual can be solved unambiguously for an aggregate expenditure function:

$$\sum_{i=1}^{N} q_i X_i^{\text{comp}}\left(\vec{\mathbf{q}}; \overline{U}\right) = M\left(\vec{\mathbf{q}}; \overline{U}\right) \tag{14.34}$$

In this case, then, aggregate dead-weight loss from taxation is also unambiguously defined as

$$L(\vec{\mathbf{t}}) = M\left(\vec{\mathbf{q}}; \overline{U}\right) - \sum_{k=2}^{N} t_k X_k^{\text{comp}}\left(\vec{\mathbf{q}}; \overline{U}\right) \tag{14.35}$$

exactly analogous to the one-consumer economy.

Unfortunately, none of the sufficient conditions is particularly compelling. Thus, it would seem more realistic to analyze second-best tax (and expenditure) problems within the context of social welfare maximization and actual general equilibria using Eqn (14.33) as the maximand, and under the assumptions of nonidentical individual preferences, a fixed distribution of lump-sum incomes (I^1, …, I^H) and unequal social welfare weights, $\partial W^*/\partial V^h$. We will adopt this approach for the remainder of the chapter.[10]

Optimal Commodity Taxation in a Many-Person Economy

As one illustration of the differences in second-best analysis between one-person (equivalent) and many-person economies, let us reconsider the optimal commodity tax problem in a many-person context, while retaining the assumption of fixed producer prices. As in the one-consumer economies, assume good 1 is the untaxed numeraire to ensure that relative prices change as tax rates are varied, with resulting losses in social welfare. Note also that with fixed producer prices, $\vec{\mathbf{p}}$, the social welfare function

$$W^*\left[V^h\left(\vec{\mathbf{q}}; \overline{I}^h\right)\right] = V\left(\vec{\mathbf{q}}; \overline{I}^1, \ldots, \overline{I}^H\right)$$

along with the pricing identities $\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$ provides a complete general equilibrium description of the economy. Production is entirely specified by the producer price vector, $\vec{\mathbf{p}}$. Market clearance is implicit, as production is perfectly elastic at the prices $\vec{\mathbf{p}}$, expanding or contracting as needed to meet the aggregate vector of consumer demands (factor supplies). Moreover, $\vec{\mathbf{q}}$ is determined by the pricing identities given $\vec{\mathbf{t}}$.

The government's problem, then, is to

$$\max_{(t_k)} W^*\left[V^h\left(\vec{\mathbf{q}}; \overline{I}^h\right)\right] = V\left(\vec{\mathbf{q}}; \overline{I}^1, \ldots, \overline{I}^H\right)$$

$$\text{s.t. } \sum_{h=1}^{H}\sum_{i=2}^{N} t_i X_{hi} = \overline{T}$$

along with the identities $\vec{\mathbf{q}} \equiv \vec{\mathbf{p}} + \vec{\mathbf{t}}; q_1 \equiv p_1 \equiv 1; t_1 \equiv 0$, where

$\overline{T}$ = the fixed amount of revenue to be collected with distorting taxes.
X_{hi} = good (factor) i demanded by (supplied by) person h, for $h = 1, \ldots, H$; $i = 1, \ldots, N$ (The X_{hi} are the actual goods demands and factor supplies and $\overline{T}$ is the actual tax revenue.).

The corresponding Lagrangian is

$$\max_{(t_k)} L = W^*\left[V^h\left(\vec{\mathbf{q}}; \overline{I}^h\right)\right] + \lambda\left(\sum_{h=1}^{H}\sum_{i=2}^{N} t_i X_{hi} - \overline{T}\right)$$

Assuming the distribution of lump-sum income is fixed, the first-order conditions are[11,12]

$$t_k: -\sum_{h=1}^{H} \frac{\partial W^*}{\partial V^h}\alpha^h X_{hk} + \lambda \sum_{h=1}^{H}\left(X_{hk} + \sum_{i=2}^{N} t_i \frac{\partial X_{hi}}{\partial q_k}\right) = 0$$
$$k = 2, \ldots, N \tag{14.36}$$

and

$$\sum_{h=1}^{H}\sum_{i=2}^{N} t_i X_{hi} = \overline{T} \tag{14.37}$$

10. Recall from Chapter 4 that Jorgenson defined a social expenditure function as the minimum aggregate income required to reach a given level of social welfare. In his model, the minimum income occurs when utilities are equal if society has any aversion to inequality. He then compares the value of the social expenditure function at different general equilibria to obtain an aggregate income measure of gain or loss. For purposes of describing optimal policy rules, however, minimizing Jorgenson's social expenditure function is not equivalent to maximizing social welfare when utilities are unequal and the distribution of income is fixed. Also, Jorgenson's social expenditure function assumes that the government can costlessly redistribute, which is best suited to a first-best environment. Harris and Wildasin described the true dual of the social welfare maximization problem in a second-best environment when redistribution is costly. Their model requires specifying the form that the government redistribution must take to hold social welfare constant, which depends on the nature of the underlying constraints. It is much more straightforward to work directly with the social welfare function. See Harris and Wildasin (1985).

11. The derivation of Eqn (14.36) employs Roy's Identity on individual's indirect utility functions, $\partial V^h/\partial q_k = -\alpha^h X_{hk}$, $k = 1, \ldots, N$. The identity follows from differentiating the consumer's indirect utility function and making use of the first-order conditions from utility maximization: $\frac{\partial V^h(X_{hi}(q,I^h))}{\partial q_k} = \sum_i \frac{\partial V^h}{\partial X_{hi}} \cdot \frac{\partial X_{hi}}{\partial q_k} = \sum_i \alpha^h q_i \frac{\partial X_{hi}}{\partial q_k} = \alpha^h \sum_i q_i \frac{\partial X_{hi}}{\partial q_k}$. But, $\sum_i q_i \frac{\partial X_{hi}}{\partial q_k} = -X_{hk}$ from differentiating the consumer's budget constraint, $\sum_i q_i X_{hi} = I^h$, with respect to q_k, generating Roy's Identity.

12. Recall that with fixed producer prices, $\partial q_i/\partial t_k = 0$, $i \neq k$.

where

α^h = the private marginal utility of income for person h.

Let $\beta^h = (\partial W^*/\partial V^h)\alpha^h$ represent the social marginal utility of income for person h, the product of the marginal social welfare weight and the private marginal utility of income. Rewrite Eqn (14.36) as

$$-\sum_{h=1}^{H}\beta^h X_{hk} + \lambda\sum_{h=1}^{H}\left(X_{hk} + \sum_{i=2}^{N} t_i\frac{\partial X_{hi}}{\partial q_k}\right) = 0 \qquad k = 2,\ldots,N \tag{14.38}$$

Equation (14.38) cannot be manipulated into simple and intuitive equal percentage change rules as in the one-consumer case, even in terms of individual's compensated demands (factors supplies). All one can say by way of a simple general interpretation is that, at the optimum, the marginal change in social welfare resulting from a change in any given tax rate must be proportional to the change in tax revenues resulting from changing the tax rate or

$$\frac{\partial W^*}{\partial t_k} = \lambda\frac{\partial T}{\partial t_k} \tag{14.39}$$

To see how the equal percentage change rule must be modified, substitute the individual consumer's Slutsky equations

$$\frac{\partial X_{hi}}{\partial q_k} = S^h_{ik} - X_{hk}\frac{\partial X_{hi}}{\partial I^h} \qquad h = 1,\ldots,H \qquad k = 1,\ldots,N \tag{14.40}$$

into Eqn (14.38) to obtain

$$-\sum_{h=1}^{H}\beta^h X_{hk} + \lambda\sum_{h=1}^{H}X_{hk} + \lambda\sum_{h=1}^{H}\sum_{i=2}^{N}t_i S^h_{ik} - \lambda\sum_{h=1}^{H}\sum_{i=2}^{N}t_i X_{hk}\frac{\partial X_{hi}}{\partial I^h} = 0 \quad k = 2,\ldots,N \tag{14.41}$$

Rearranging terms, dividing through by $\lambda\sum_{h=1}^{H}X_{hk} = \lambda X_k$, and noting that $S^h_{ik} = S^h_{ki}$, yields

$$\frac{\sum_{i=1}^{H}\sum_{i=2}^{N}t_i S^h_{ki}}{X_k} = -1 + \frac{\frac{1}{\lambda}\sum_{h=1}^{H}\beta^h X_{hk}}{X_k} + \frac{\sum_{i=2}^{N}\sum_{h=1}^{H}t_i X_{hk}\frac{\partial X_{hi}}{\partial i^h}}{X_k} \tag{14.42}$$

Martin Feldstein defined the *distributional coefficient for good k* as

$$\lambda^k = \sum_{h=1}^{H}\beta^h X_{hk}/X_k$$

The distributional coefficient is a weighted average of the individual social marginal utilities of income, with the weights equal to the proportion of total X_k consumed by each person. Using Feldstein's distributional coefficient, Eqn (14.42) becomes

$$\frac{\sum_{h=1}^{H}\sum_{i=2}^{N}t_i S^h_{ki}}{X_k} = -1 + \frac{\lambda^k}{\lambda} + \frac{\sum_{h=1}^{H}\sum_{i=2}^{N}t_i\frac{\partial X_{hi}}{\partial I^h}X_{hk}}{X_k} \qquad k = 2,\ldots,N \tag{14.43}$$

The left-hand side of Eqn (14.43) gives the percentage change in the aggregate compensated demand for good k (approximately), but the right-hand side (RHS) is no longer independent of k. Rather, the percentage changes depend in a complicated manner on Feldstein's distributional coefficients and the change in tax revenue in response to changes in the pattern of lump-sum incomes. Furthermore, the RHS cannot readily be divided into two distinct sets of terms, with one set containing all relevant efficiency considerations and the second containing all relevant distributional information.

One can shed some additional light on the pattern of optimal taxes by considering changes in actual demands, even though these changes cannot be described in a simple way either. From the individual Slutsky equations,

$$S^h_{ki} = \frac{\partial X_{hk}}{\partial q_i} + X_{hi}\frac{\partial X_{hk}}{\partial I^h} \tag{14.44}$$

Substituting for the S^h_{ki} in Eqn (14.43) and rearranging terms,

$$\frac{\sum_{h=1}^{H}\sum_{i=2}^{N}t_i\frac{\partial X_{hk}}{\partial q_i}}{X_k} = -1 + \frac{\lambda^k}{\lambda} + \frac{\sum_{h=1}^{H}\sum_{i=2}^{N}t_i\frac{\partial X_{hi}}{\partial I^h}X_{hk}}{X_k} - \frac{\sum_{h=1}^{H}\left(\sum_{i=2}^{N}t_i X_{hi}\right)\frac{\partial X_{hk}}{\partial I^h}}{X_k} \qquad k = 2,\ldots,N \tag{14.45}$$

$$\frac{\sum_{h=1}^{H}\sum_{i=2}^{N}t_i\frac{\partial X_{hk}}{\partial q_i}}{X_k} = -1 + \frac{\lambda^k}{\lambda} + \frac{\sum_{h=1}^{H}\sum_{i=2}^{N}t_i\frac{\partial X_{hi}}{\partial I^h}X_{hk}}{X_k} - \sum_{h=1}^{N}\left(\frac{\sum_{i=2}^{N}t_i X_{hi}}{I^h}\cdot\frac{\partial X_{hk}}{\partial I^h}\cdot\frac{I^h}{X_{hk}}\right) \qquad k = 2,\ldots,N \tag{14.46}$$

Equation (14.46) says that the actual percentage changes in demand (factor supply) resulting from the optimal pattern of commodity taxes should be greater:

1. The lower its distributional coefficient λ^k or the more it is demanded by people with low social marginal utilities of income ($\Delta X_k/X_k$ is expected to be negative for goods, and λ is positive). Presumably, the people with low λ^k are the rich. If so, the rule says that, other things equal, taxes should be the heaviest on those goods consumed most heavily by the rich.

2. The more it is demanded by people whose total taxes change the least as lump-sum income changes.
3. The more it is demanded by people for whom, other things equal, the product of the fraction of income paid as taxes and the income elasticity of demand for the good is the highest.

Unfortunately, there is no clear presumption as to whom the people referred to in items 2 and 3 might be, so that rewriting the first-order conditions in terms of actual demand changes still fails to provide any really clear intuitive feel for the optimal pattern of taxation.

A Covariance Interpretation of Optimal Taxation

Peter Diamond provided an ingenious interpretation of these rules that does give one a better intuitive appreciation of the tax rules (Diamond, 1975). Suppose the government, in addition to the commodity taxes, has the ability to offer a single head or poll subsidy of equal value to all consumers. Although this is admittedly a lump-sum subsidy, it is not the sophisticated variable subsidy necessary to satisfy the interpersonal equity conditions of the first-best theory. With the additional head subsidy, the government's problem becomes

$$\max_{(\vec{\mathbf{t}},I)} \; V(\vec{\mathbf{q}}; I^1, \ldots, I^H)$$

$$\text{s.t.} \quad \sum_{h=1}^{H}\sum_{i=2}^{N} t_i X_{hi} = \overline{T} + H \cdot I$$

where

I = the equal per-person subsidy.

The first-order conditions with respect to the t_k are obviously unchanged by the presence of the subsidy. Reproducing Eqn (14.41),

$$-\sum_{h=1}^{H}\beta^h X_{hk} + \lambda\sum_{h=1}^{H} X_{hk} + \lambda\sum_{h=1}^{H}\sum_{i=2}^{N} t_i S_{ik}^h - \lambda\sum_{h=1}^{H}\sum_{i=2}^{N} t_i X_{hk}\frac{\partial X_{hi}}{\partial I^h} = 0 \quad k = 2, \ldots, N \tag{14.47}$$

The first-order condition with respect to the head tax, I, is

$$\sum_{h=1}^{H}\beta^h + \lambda\left(\sum_{h=1}^{H}\sum_{i=2}^{N} t_i\frac{\partial X_{hi}}{\partial I} - H\right) = 0 \tag{14.48}$$

with $\partial I = \partial I^h$, all $h = 1, \ldots, H$. Diamond then defines

$$\gamma^h = \beta^h + \lambda\sum_{i=2}^{N} t_i\frac{\partial X_{hi}}{\partial I^h} \tag{14.49}$$

as the full social marginal utility of income for person h, consisting of the conventional direct increase in social utility when I^h increases, the β^h term, plus the social marginal utility of the increased tax revenues when I^h increases, equal to $\lambda\sum_{i=2}^{N} t_i(\partial X_{hi}/\partial I^h)$. With γ^h defined in this manner, the first-order conditions, Eqn (14.47), can be rewritten as

$$\lambda\sum_{i=2}^{N}\sum_{h=1}^{H} t_i S_{ik}^h = \sum_{h=1}^{H}(\gamma^h - \lambda)\cdot X_{hk} \quad k = 2, \ldots, N \tag{14.50}$$

Furthermore, Eqn (14.48) becomes simply

$$\lambda H = \sum_{h=1}^{H}\gamma^h \tag{14.51}$$

or

$$\lambda = \frac{\sum_{h=1}^{H}\gamma^h}{H} \tag{14.52}$$

Thus, λ can be interpreted as the average full social marginal utility of income given that the government employs an optimal head subsidy. Furthermore, once λ is expressed in this form, the first-order conditions, Eqn (14.50), have a simple covariance interpretation. To see this, divide Eqn (14.50) by $\lambda X_k = \lambda\sum_{h=1}^{H} X_{hk}$ (and note that $S_{ik}^h = S_{ki}^h$) to obtain

$$\frac{\sum_{h=1}^{H}\sum_{i=2}^{N} t_i S_{ki}^h}{X_k} = \frac{\sum_{h=1}^{H}(\gamma^h - \lambda)X_{hk}}{\lambda X_k} \quad k = 2, \ldots, N \tag{14.53}$$

But, from Eqn (14.51), $\sum_{h=1}^{H}(\gamma^h - \lambda) = 0$. Hence, $\sum_h(\gamma^h - \lambda)\cdot\overline{X}_k = 0$, where

$$\overline{X}_k = \frac{\sum_{h=1}^{H} X_{hk}}{H}$$

so that Eqn (14.53) can be rewritten as

$$\frac{\sum_{i=1}^{H}\sum_{i=2}^{H} t_i S_{ki}^h}{X_k} = \frac{\sum_{h=1}^{H}(\gamma^h - \lambda)\left(X_{hk} - \overline{X}_k\right)}{H\lambda\overline{X}_k} \quad k = 2, \ldots, N \tag{14.54}$$

Equation (14.54) says that the aggregate percentage change in the compensated demand (supply) of good (factor) k should be proportional to the covariance between the full marginal social utility of income and the consumption (supply) of good (factor) k. This is the simplest interpretation of the many-person optimal tax rule to date (although it requires the simultaneous imposition of a uniform head subsidy/tax).

A Two-Class Tax Rule

Defining an optimal per-person income subsidy and the full social marginal utility of income yields some additional intuition into the pattern of optimal taxes. Consider again the first-order conditions, Eqn (14.50), with λ interpreted as the average full social marginal utility of income given an optimal head subsidy. Multiply each equation by t_k and sum over $k = 1, \ldots, N$ to obtain[13]

$$\sum_{h=1}^{H}\left[\left(\gamma^h - \lambda\right)\sum_{k=1}^{N} t_k X_{hk}\right] = \lambda \sum_{h=1}^{H}\sum_{i=1}^{N}\sum_{k=1}^{N} t_i S_{ik}^h t_k \quad (14.55)$$

Because S_{ik}^h is negative semidefinite,

$$\sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{k=1}^{N} t_i S_{ik}^h t_k \leq 0$$

Therefore,

$$\sum_{h=1}^{H}\left[\left(\gamma^h - \lambda\right)\sum_{k=1}^{N} t_k X_{hk}\right] \leq 0 \quad (14.56)$$

Suppose the government is willing to think of the H consumers as divided into two subsets, the rich and the poor, such that all rich people are identical and all poor people are identical (equal preferences and equal full social marginal utilities of income). Let there be R rich people each with full social marginal utility γ^R, and $(H - R)$ poor people each with full social marginal utility of income γ^P, such that $\gamma^P > \gamma^R$ and $\gamma^P > \lambda$, where λ is the average full social marginal utility of income over all H people.[14] With an optimal head subsidy (Eqn (14.51) satisfied),

$$\sum_{h=1}^{H}\gamma^h = \lambda H = R\cdot\gamma^R + (H - R)\gamma^P \quad (14.57)$$

Substituting Eqn (14.57) into Eqn (14.56) yields

$$R(\gamma^R - \lambda)\sum_{k=1}^{N} t_k X_{Rk} + (H - R)(\gamma^P - \lambda)\sum_{k=1}^{N} t_k X_{Pk} \leq 0 \quad (14.58)$$

But, from Eqn (14.57),

$$[R + (H - R)]\lambda = R\gamma^R + (H - R)\gamma^P \quad (14.59)$$

Rearranging terms,

$$R(\gamma^R - \lambda) = -(H - R)(\gamma^P - \lambda) \quad (14.60)$$

Substituting for $R(\gamma^R - \lambda)$ in Eqn (14.58) yields

$$-(H - R)(\gamma^P - \lambda)\sum_{k=1}^{N} t_k X_{Rk} + (H - R)(\gamma^P - \lambda)\sum_{k=1}^{N} t_k X_{Pk} \leq 0 \quad (14.61)$$

Rearranging terms,

$$(H - R)(\gamma^P - \lambda)\left(\sum_{k=1}^{N} t_k X_{Pk} - \sum_{k=1}^{N} t_k X_{Rk}\right) \leq 0 \quad (14.62)$$

Hence, assuming $(\gamma^P - \lambda) > 0$ implies that $\sum_{k=1}^{N} t_k X_{Rk} \geq \sum_{k=1}^{N} t_k X_{Pk}$, or that the optimal pattern of commodity taxes should, in general, collect more taxes from the rich than the poor.

This result is certainly consistent with one's intuitive sense of the effect of social welfare considerations on the optimal pattern of commodity taxes. Nonetheless, Eqn (14.56) does not necessarily yield such simple guidelines when there are more than two classes of people. Also, Eqn (14.62) is more or less compelling depending on how one defines the poor. If the "poor" refers to those in poverty, roughly 11% of the population in the United States, then Eqn (14.62) would be satisfied by almost any tax or set of taxes, not just optimal taxes. If, however, the "poor" refers to those with incomes below the median and $(\gamma^P - \lambda)$ is assumed to be positive for them, then Eqn (14.62) suggests that the optimal taxes might be progressive with respect to the two groups.

US Commodity Taxes: How Far from Optimal?

Balcer et al. applied a many-person, fixed-producer-price model to US data to get a feel for how optimal commodity tax rates would vary with the government's revenue needs and society's aversion to inequality.[15] Using data from the 1972 and 1973 Consumer Expenditure Surveys, they calculated expenditures for nine commodity groups for each of 10 income classes. The specific model they used to calculate optimal commodity taxes for these data had the following features:

1. Preferences given by the Stone–Geary utility function $U = \sum_{i=1}^{9}(X_i - \gamma_i)^{\beta_i}$, where the γ_i are the subsistence quantities of each good (the minimum quantities above which utility is positive) and the β_i are the marginal budget shares of each good. The γ_i are assumed to be the expenditures of the poorest income class calculated at the net-of-tax producer prices. The β_i vary by income class. The poorest income class is assumed to receive

13. With $t_1 = 0$, k or i can be summed from 1 or 2 to N.
14. Refer to Eqn (14.52) and its derivation.
15. Balcer et al. (1983), Chapter 13.

subsidies that just place them at the subsistence level, so that class is dropped from the analysis. Labor is in fixed supply, and there is no saving.

2. An Atkinson social welfare function

$$W = \frac{1}{(1-e)} \sum_{h=1}^{H} \left[V^h(q_i; I^h)\right]^{(1-e)}$$

where

V^h = the indirect utility function for income class h

$q_i = 1 + t_i$, the gross-of-tax consumer price for good i (producer prices are set equal to unity, so that expenditures equal quantities at the net-of-tax prices);

e = society's aversion to inequality, ranging from 0 (utilitarian) to 2. They believe that e is likely to be in the neighborhood of 0.5 for the United States with an outside range of 0.25−0.75, much as Harberger conjectured (see the discussion in Chapter 4). Like Harberger, they believe that the United States does not have much aversion to inequality.

3. A government budget constraint of the form $\sum_{h=1}^{H}\sum_{i=1}^{N} t_i X_{hi} = R$, where R represents the revenue needs of the government. R varied from −5% to 30% of total disposable income in their exercises. The actual revenues collected from US sales and excise taxes at the time were slightly in excess of 4% of disposable income.

The optimal commodity taxes are those that maximize W subject to the government budget constraint. The exercise produced results that were generally consistent with the theory and gave fairly high marks to the actual tax rates on the nine commodity groups. Among their more interesting results are the following:

1. For the baseline case of $e = 0.5$ and R slightly in excess of 4%, the optimal tax rates range from −10.8% (housing) to 11.4% (recreation). They vary directly with the ratio of marginal budget shares of the rich to the poor, as predicted by the theory. The variation in rates increases as e and R increase, also as predicted by the theory.
2. Changing the US tax rates from their (then) current values to the optimal values that generate the same revenue would produce only a small increase in social welfare (asserted by the authors on p. 292, but with no data given). It would also have a negligible effect on the Gini coefficient. The reductions in the Gini range from 0.68% to 2.96% for the values of e (positive) and R tested. The reason for the small distributional gain is that only one tax deviated substantially from the optimal rate, the tax on gasoline (39.6% vs. a baseline optimal rate of 8.8%). (The high gasoline tax could be justified as a benefits-received tax since its revenues are typically earmarked to highway funds.)
3. They also considered the welfare loss of moving to uniform taxes from the optimal tax rates, with the loss defined as the amount that the government would have to lower its revenue to maintain social welfare at its value with the optimal rates. The welfare cost is negligible, only 0.17% of disposable income for the baseline case and never higher than 2.28% across all values of R and e. (The welfare cost is zero in the utilitarian case, $e = 0$, since that is equivalent to a one-consumer economy and the Stone−Geary utility function satisfies the sufficient condition for uniform taxation to be optimal in the one-consumer model.)

MANY-PERSON ECONOMY WITH GENERAL TECHNOLOGY

Synthesizing the separate analyses of a one-person economy with general technology and the many-person economy with linear technology (constant producer prices) is relatively straightforward, especially under the assumption of CRS production.

Let us begin by considering the optimal commodity tax problem. We saw that assuming CRS in the context of a one-consumer economy generates the same optimal tax rules that result when production technology is characterized by fixed producer prices. The key to this result is that there can never be pure economic profits or losses under CRS and perfect competition, so that the value of the general equilibrium profit function $\pi(\vec{\mathbf{p}})$ is identically zero for all values of the producer price vector $\vec{\mathbf{p}}$.

The same correspondence exists in the many-person economy. As long as we assume CRS, the original distribution of lump-sum incomes (I^1, …, I^H) remains unchanged as producer prices vary in response to taxation. Hence, the many-person optimal tax rule is identical to its linear technology counterpart. In fact, Diamond used a general technology CRS model to generate the many-person optimal tax rules.

A model appropriate for analyzing second-best tax (and expenditure) problems in a many-person, general technology economy has four components: the social welfare function, consumer preferences, a general production technology, and market clearance.

Social Welfare and Preferences

The object of government policy is to maximize a social welfare function of the form

$$W\left[V^h(\vec{\mathbf{q}}; I^h)\right] = V(\vec{\mathbf{q}}; I^1, \ldots, I^H)$$

specified in terms of consumer prices, exactly as in the many-person, linear technology case. (We drop the * on W here.)

Production Technology

Production must be specified in terms of prices and actual general equilibria to be compatible with social welfare. The specification must also be flexible enough to allow for various kinds of technologies. But the general technology production can no longer be specified by means of the generalized profit function, $\pi(\vec{\mathbf{p}})$, as in the one-consumer economy, because social welfare is not measured in terms of lump-sum income. Instead, the natural choice is to return to an implicit aggregate production frontier of the form $F(\vec{\mathbf{Y}}) = 0$, as in the first-best analysis, where $\vec{\mathbf{Y}}$ = the vector of aggregate goods supplies (factor demands). Then, replace the quantities $\vec{\mathbf{Y}}$ with the general equilibrium market supply (input demand) functions $Y_i = Y_i(\vec{\mathbf{p}})$, $i = 1, \ldots, N$ (the same functions that would result from a social planner maximizing aggregate profits at given competitive prices). The resulting function, $F[\vec{\mathbf{Y}}(\vec{\mathbf{p}})] = 0$, which is called the production-price frontier, specifies all relevant production parameters assuming competitive market behavior.

Market Clearance

General technology requires explicit market clearance equations of the form

$$\sum_{h=1}^{H} X_{hi}\left(\vec{\mathbf{p}} + \vec{\mathbf{t}}, I^h\right) = Y^i\left(\vec{\mathbf{p}}\right) \quad i = 1, \ldots, N \qquad (14.63)$$

to solve for the vector of producer prices given a vector of tax rates. All N market clearance equations apply because the model describes an actual general equilibrium, not a compensated general equilibrium. The pricing identities $\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$ then solve for the vector of consumer prices.

The Model

Thus, a full general equilibrium model useful for analyzing any problem in the second-best theory of taxation can be represented as

$$\begin{aligned} &\max_{(\vec{q}, \vec{t}, \vec{p})} \; W\left[V^h\left(\vec{\mathbf{q}}; \bar{I}^h\right)\right] \\ &\text{s.t.} \quad F\left[\vec{\mathbf{Y}}\left(\vec{\mathbf{p}}\right)\right] = 0 \\ &\qquad \sum_{h=1}^{H} X_{hi}\left(\vec{\mathbf{q}}; I^h\right) = Y_i\left(\vec{\mathbf{p}}\right) \quad 1, \ldots, N \\ &\qquad q_i = p_i + t_i \quad i = 2, \ldots, N \\ &\qquad q_1 \equiv p_1 \equiv 1 \quad t_1 = 0 \end{aligned}$$

As always, setting $t_1 = 0$ ensures that the tax vector $\vec{\mathbf{t}}$ changes the vector of relative consumer and producer prices and thereby generates distortions.

The model can be greatly simplified by incorporating market clearance directly into the production frontier and thinking of the government as solving directly for the vector of consumer prices, $\vec{\mathbf{q}}$, rather than the vector of taxes, $\vec{\mathbf{t}}$ as follows:

$$\begin{aligned} &\max_{(q)} \; W\left[V^h\left(\vec{\mathbf{q}}; I^h\right)\right] \\ &\text{s.t.} \quad F\left[\sum_{h=1}^{H} X_{hi}\left(\vec{\mathbf{q}}; I^h\right)\right] = 0 \end{aligned}$$

The vector of producer prices $\vec{\mathbf{p}}$ can then be determined through the market clearance equations, after which the $(N-1)$ optimal tax rates are given by the pricing identities $t_i = q_i - p_i$, $i = 2, \ldots, N$.

Walras' Law and the Government Budget Constraint

The final point is that there is no need to include the government's budget constraint,

$$\sum_{h=1}^{H} \sum_{i=2}^{N} t_i X_{hi}\left(\vec{\mathbf{q}}; I^h\right) = \bar{T}$$

explicitly in the model because of Walras' law. The model describes an actual market general equilibrium, for which Walras' law can have either of two interpretations:

1. *The common interpretation*: If each economic agent is on its budget constraint (firms are profit maximizing) and all but one market is in equilibrium, the final market must also be in equilibrium.
2. *An alternative interpretation*: If all but one economic agent are satisfying their budget constraints and all markets are in equilibrium, then the last economic agent must also be on its budget constraint. This is the interpretation that allows us to exclude the government's budget constraint since the model (a) explicitly posits market clearance in all markets, (b) implicitly assumes all consumers are on their budget constraints as a prerequisite for defining their indirect utility functions, and (c) implicitly assumes all producers are maximizing profits when substituting the general equilibrium supply (input demand) functions $\vec{\mathbf{Y}}(\vec{\mathbf{p}})$ into the aggregate production frontier $F(\vec{\mathbf{Y}}) = 0$.

This is the model specification used by Diamond to generate many-person optimal taxes rules identical to Eqns (14.50) and (14.52).[16]

Optimal Taxation

To see that the assumption of general technology makes no difference as long as the technology exhibits CRS, consider

16. The model is not identical to Diamond's model, since he included a Samuelson nonexclusive public good and assumed all consumers had identical initial endowments of lump-sum income.

the first-order conditions of the model with respect to q_k and an equal head subsidy, I:

$$\sum_{h=1}^{H} \frac{\partial W}{\partial V^h} \frac{\partial V^h}{\partial q_k} = \lambda \sum_{i=1}^{N} \sum_{h=1}^{H} F_i \frac{\partial X_{hi}}{\partial q_k} \quad k = 2, \ldots, N \tag{14.64}$$

$$\sum_{h=1}^{H} \frac{\partial W}{\partial V^h} \frac{\partial V^h}{\partial I} = \lambda \sum_{i=1}^{N} \sum_{h=1}^{H} F_i \frac{\partial X_{hi}}{\partial I} \tag{14.65}$$

Equation (14.64) implicitly embodies the assumption of CRS production because the initial distribution of lump-sum income, $(I^1,\ldots,I^H)$, is assumed unchanged by a marginal change in the kth consumer price.

From Roy's Identity on indirect utility functions, the definition of marginal social utility β^h, and the assumption of profit maximization with $p_1 \equiv 1$, Eqn (14.64) can be written as[17]

$$-\sum_{h=1}^{H} \beta^h X_{hk} = \lambda \sum_{h=1}^{H} \sum_{i=1}^{N} p_i \frac{\partial X_{hi}}{\partial q_k} \quad k = 2, \ldots, N \tag{14.66}$$

But $p_i = q_i - t_i$, for $i = 1, \ldots, N$. Hence,

$$-\sum_{h=1}^{H} \beta^h X_{hk} = \lambda \sum_{h=1}^{H} \sum_{i=1}^{N} \left(q_i \frac{\partial X_{hi}}{\partial q_k} - t_i \frac{\partial X_{hi}}{\partial q_k} \right) \quad k = 2, \ldots, N \tag{14.67}$$

Further, if consumers are on their budget constraints,

$$\sum_{i=1}^{N} q_i \frac{\partial X_{hi}}{\partial q_k} = -X_{hk} \quad h = 1, \ldots, H \tag{14.68}$$

Therefore,

$$-\sum_{h=1}^{H} \beta^h X_{hk} = \lambda \sum_{h=1}^{H} \left(-X_{hk} - \sum_{i=1}^{N} t_i \frac{\partial X_{hi}}{\partial q_k} \right) \quad k = 2, \ldots, N \tag{14.69}$$

which is identical to Eqn (14.38).[18]

Turning to the optimal head tax, make use of profit maximization, the definition of marginal social utility, and the definitional relationships among prices and taxes, to rewrite Eqn (14.65) as

$$\sum_{h=1}^{H} \beta^h = \lambda \sum_{h=1}^{H} \sum_{i=1}^{N} (q_i - t_i) \frac{\partial X_{hi}}{\partial I} \tag{14.70}$$

If consumers are on their budget constraints,

$$\sum_{i=1}^{N} q_i \frac{\partial X_{hi}}{\partial I} = 1 \quad h = 1, \ldots, H \tag{14.71}$$

Hence,

$$\sum_{h=1}^{H} \beta^h = \lambda H - \lambda \sum_{h=1}^{H} \sum_{i=1}^{N} t_i \frac{\partial X_{hi}}{\partial I} \tag{14.72}$$

But

$$\gamma^h = \beta^h + \lambda \sum_{i=1}^{N} t_i \frac{\partial X_{hi}}{\partial I}$$

the full social marginal utility of income. Therefore,

$$\lambda \sum_{h=1}^{H} \frac{\gamma^h}{H} \tag{14.73}$$

the average full social marginal utility of income, as in Eqn (14.52). Consequently, the many-person optimal tax rules continue to have a simple covariance interpretation.

THE SOCIAL WELFARE IMPLICATIONS OF ANY GIVEN CHANGE IN TAXES

Once the optimal commodity tax problem had been fully developed by Diamond and Mirrless in the late 1960s, public sector economists turned their attention to more realistic forms of restricted taxation. This opened up two new major lines of research in the 1970s. One group of economists adopted the basic model for optimal commodity taxation and attempted to develop theorems on optimal changes (or levels) of taxes for a subset of the goods and factors (e.g., Dixit, Guesnerie, and Hatta).[19] A second group, following the lead of James Mirrlees and Ray Fair in 1971, concentrated specifically on optimal income taxation (e.g., Mirrlees, Fair, Sheshinski, Atkinson, Stiglitz, Sadka, Stern, and Seade).[20] The chapter concludes with a general example representative of the first line of research. Chapter 15 discusses optimal income taxation.

The general method for analyzing restricted tax changes can been seen by considering the social welfare implications of a marginal change in a single tax, or of substituting one vector of tax rates for another, equal-revenue vector of rates in the context of a many-person, general technology economy.[21] Begin by totally differentiating the social welfare function $W^* = W[V^h(\vec{\mathbf{q}}; I^h)]$ with respect to prices and income. Using Roy's Identity and the definition of social marginal utility of income β^h,

17. From profit maximization $F_i/F_1 = p_i/p_1$, but $p_1 \equiv 1$ and F can be scaled such that $F_1 = 1$, so that $F_i = p_i$, $i = 2, \ldots, N$.
18. $\sum_{i=1}^{N} t_i \frac{\partial X_{hi}}{\partial q_k} = \sum_{i-2}^{N} t_i \frac{\partial X_{hi}}{\partial q_k}$, with $t_1 = 0$.
19. The seminal articles were Dixit (1975), Dixit and Munk (1977), Guesnerie (1977), Guesnerie (1979), and Hatta (1977).
20. The seminal articles were Mirrlees (1971) (*the* seminal article), Mirrlees (1976), Fair (1971), Sheshinski (1972), Atkinson (1973), Atkinson and Stiglitz (1976), Sadka (1976), Stern (1976), Seade (1977), Bradford and Rosen, 1976.
21. The analysis in this section draws heavily from Boadway (1976).

$$dW = -\sum_{h=1}^{H}\sum_{i=1}^{N}\frac{\partial W}{\partial V^h}\alpha^h X_{hi}dqi + \sum_{h=1}^{H}\frac{\partial W}{\partial V^h}\alpha^h dI^h \quad (14.74)$$

$$dW = -\sum_{h=1}^{H}\sum_{i=1}^{N}\beta^h X_{hi}dq_i + \sum_{h=1}^{H}\beta^h dI^h \quad (14.75)$$

Next, totally differentiate the production-price frontier $F(\sum_{h=1}^{H}X_{hi}) = 0$, in which the market clearance equations have been used to substitute consumers' demands and factor supplies for the production aggregates Y_i:

$$\sum_{i=1}^{N}F_i\sum_{h=1}^{H}dX_{hi} = 0 \quad (14.76)$$

Assuming perfect competition, $p_1 \equiv 1$, and that the identity of person h is irrelevant to production, Eqn (14.76) becomes

$$\sum_{i=1}^{N}p_i\sum_{h=1}^{H}dX_{hi} = 0 \quad (14.77)$$

But, $q_i = p_i + t_i$, for $i = 1, \ldots, N$. Multiplying each price by dX_{hi}, and summing over all goods and people yields

$$\sum_{i=1}^{N}\sum_{h=1}^{H}q_i dX_{hi} = \sum_{i=1}^{N}\sum_{h=1}^{H}(p_i + t_i)dX_{hi} \quad (14.78)$$

which, from Eqn (14.78), becomes

$$\sum_{i=1}^{N}\sum_{h=1}^{H}q_i dX_{hi} = \sum_{i=1}^{H}\sum_{h=1}^{H}t_i dX_{hi} \quad (14.79)$$

Next, totally differentiate each consumer's budget constraint and sum over all consumers to obtain

$$\sum_{h=1}^{H}dI^h = \sum_{h=1}^{H}\sum_{i=1}^{N}q_i dX_{hi} + \sum_{h=1}^{H}\sum_{i=1}^{N}X_{hi}dq_i \quad (14.80)$$

Combining Eqns (14.79) and (14.80) yields

$$\sum_{h=1}^{H}dI^h - \sum_{h=1}^{H}\sum_{i=1}^{N}X_{hi}dq_i = \sum_{h=1}^{H}\sum_{i=1}^{N}t_i dX_{hi} \quad (14.81)$$

Thus, Eqn (14.75) can be written as

$$dW = -\sum_{h=1}^{H}\sum_{i=1}^{N}\beta^h X_{hi}dq_i + \sum_{h=1}^{H}\beta^h dI^h - \sum_{h=1}^{H}dI^h + \sum_{h=1}^{H}\sum_{i=1}^{N}X_{hi}dq_i + \sum_{i=1}^{N}\sum_{h=1}^{H}t_i dX_{hi} \quad (14.82)$$

Finally, the dX_{hi} in the last term of Eqn (14.82) can be eliminated by noting that

$$X_{hi} = X_{hi}(\vec{\mathbf{q}}; I^h) \quad h = 1, \ldots, H; \quad i = 1, \ldots N \quad (14.83)$$

Totally differentiating Eqn (14.83) yields

$$dX_{hi} = \sum_{j=1}^{N}\frac{\partial X_{hi}}{\partial q_j}dq_j + \frac{\partial X_{hi}}{\partial I^h}di^h \quad h = 1, \ldots H \quad i = 1, \ldots, N \quad (14.84)$$

Substituting Eqn (14.84) into Eqn (14.82) and combining terms yields

$$dW = \sum_{h=1}^{H}\left(\beta^h - 1 + \sum_{i=1}^{N}\sum_{h=1}^{H}t_i\frac{\partial X_{hi}}{\partial I^h}\right)dI^h + \sum_{h=1}^{H}(1-\beta^h)\sum_{i=1}^{N}X_{hi}dq_i + \sum_{h=1}^{H}\sum_{i=1}^{N}t_i\sum_{j=1}^{N}\frac{\partial X_{hi}}{\partial q_j}dq_j \quad (14.85)$$

Equation (14.85) highlights the importance of CRS in the second-best analysis. With general technology, pure profits or losses can occur in production, thereby changing the pattern of lump-sum incomes $(I^1, \ldots, I^H)$ received by the consumers. As indicated by the first term in Eqn (14.85), the government would then have to keep track of these changes and their subsequent effects on social welfare. With CRS, however, the first term can be ignored since pure profits and losses are zero and the vector of lump-sum income remains unchanged.

Even with CRS, however, it is clear that production derivatives affect second-best decision rules, in general, even if they do not do so in the optimal tax problem. The change in the vector of consumer prices, $\vec{\mathbf{q}}$, in Eqn (14.85) is determined by the combined interaction of general equilibrium demand and supply schedules. To see this explicitly, ignore changes in lump-sum income and use market clearance to express the change in welfare in terms of changes in taxes rather than prices, exactly as we did for the one-consumer, general technology case. Totally differentiating the market clearance equations,

$$\sum_{h=1}^{H}X_{hi}(\vec{\mathbf{q}}; \vec{I}^h) = Y_i(\vec{\mathbf{p}}) = Y_i(\vec{\mathbf{q}} - \vec{\mathbf{t}}) \quad i = 1, \ldots N$$

yields

$$\sum_{h=1}^{H}\sum_{j=1}^{N}\frac{\partial X_{hi}}{\partial q_j}dq_j = \sum_{j=1}^{N}\frac{\partial Y_i}{\partial p_j}(dq_j - dt_j) \quad i = 1, \ldots, N \quad (14.86)$$

Solving for $d\vec{\mathbf{q}}$ and expressing the N equations, Eqn (14.86), in vector notation yields

$$dq = E^{-1}\left(-\frac{\partial Y}{\partial p}\right)dt \quad (14.87)$$

where $E = \left(\frac{\partial X}{\partial q} - \frac{\partial Y}{\partial p}\right)$ in vector notation.[22]

22. Each element ij in X is the partial derivative of the aggregate X_i with respect to q_j, the sum of the H individual derivatives.

Finally, substitute Eqn (14.87) into Eqn (14.85), with $dI^h \equiv 0$, to obtain, in vector notation,

$$dW = \left[-\Big((1-\beta)'X\Big) - t'\frac{\partial X}{\partial q}\right] E^{-1}\frac{\partial Y}{\partial p}dt \qquad (14.88)$$

where

$\beta = \begin{bmatrix} \beta^1 \\ \vdots \\ \beta^H \end{bmatrix}$, an $(H \times 1)$ column vector of marginal social utilities of income.

$X = [X_{hi}]$, an $(H \times N)$ matrix of individual consumer demands and factor supplies.

$1 =$ an $(H \times 1)$ unit column vector.

Equation (14.88) is the fundamental equation for evaluating tax changes in a many-consumer economy with CRS general production technology. By inspection, the supply responses $(\partial Y/\partial p)$ affect the change in social welfare.

Equation (14.88) can also be compared directly with the results from a one-consumer equivalent economy. Equation (14.17), reproduced here as Eqn (14.89), calculated the change in loss as

$$dL = (t')\big(M_{ij}\big)E^{-1}\frac{\partial Y}{\partial p}dt \qquad (14.89)$$

The second term in Eqn (14.88) is very close to Eqn (14.89) but not identical. A trivial difference is the minus sign, resulting from the fact that $dW = -dL$. More importantly, the demand derivatives $(\partial X/\partial q)$ in Eqn (14.89) are the compensated Slutsky terms, not the actual demand derivatives, reflecting the fact that Eqn (14.89) derives from a conceptual compensation experiment that is not particularly meaningful in a many-person environment. In practical applications, however, it may prove useful to think of the change in social welfare resulting from any change in tax rates as a linear combination of social welfare considerations and dead-weight efficiency loss, with the former embodied in the first term of Eqn (14.88) and the latter in the second term. This interpretation maintains the dichotomy between equity and efficiency that exists in first-best analysis, but it can only be viewed here as a rough "interpretative" approximation. Whether it is useful or not depends on the particular problem under consideration. We saw, for example, that equity and efficiency terms are tightly intertwined in the many-person optimal commodity tax rules. But it could be more compelling for simple tax change problems, such as in the Corlett and Hague analysis. The welfare effects of such changes can be evaluated directly by Eqn (14.88),[23] once the equal-revenue pattern of tax changes, $d\vec{\mathbf{t}}$, has been determined.

23. Alternatively, Eqn (14.85) with nonconstant returns to scale production. In this case, the vector of income changes $dI^{\vec{\mathbf{h}}}$ would have to be specified and incorporated into the total differential of the market clearance equations, Eqn (14.83).

REFERENCES

Atkinson, A., 1973. How progressive should income tax be? In: Parkin, M. (Ed.), Essays in Modern Economics. Longman Group, Ltd, London.

Atkinson, A., Stiglitz, J., July/August 1976. The design of tax structure: direct vs. indirect taxation. Journal of Public Economics 6 (1-2), 55–77.

Balcer, Y., Garfinkel, I., Krynski, K., Sadka, E., 1983. Income redistribution and the structure of indirect taxation. In: Helpman, E., Razin, A., Sadka, E. (Eds.), Social Policy Evaluation: An Economic Perspective. Academic Press, New York.

Boadway, R., July 1975. Cost-benefit rules and general equilibrium. Review of Economic Studies 42 (3), 361–374.

Boadway, R., November 1976. Integrating equity and efficiency in applied welfare economics. Quarterly Journal of Economics 90 (4), 541–556.

Bradford, D., Rosen, H., May 1976. The optimal taxation of commodities and income. American Economic Association Papers and Proceedings 66 (2), 94–101.

Diamond, P.A., February 1975. A many person Ramsey tax rule. Journal of Public Economics 4 (1).

Dixit, A., February 1975. Welfare effects of tax and price changes. Journal of Public Economics 4 (2), 103–123.

Dixit, A., Munk, K., August 1977. Welfare effects of tax and price changes: a correction. Journal of Public Economics 123 (1), 103–107.

Fair, R., November 1971. The optimal distribution of income. Quarterly Journal of Economics 85 (4), 551–579.

Feldstein, M., March 1972. Distributional equity and the optimal structure of public prices. American Economic Review 62 (1), 32–36.

Green, J., November 1975. Two models of optimal pricing and taxation. Oxford Economic Papers 27 (3), 352–382.

Guesnerie, R., April 1977. On the direction of tax Reform. Journal of Public Economics 7 (2), 179–202.

Guesnerie, R., March 1979. Financing public goods with commodity taxes: a tax reform viewpoint. Econometrica 47 (2), 393–421.

Harris, R., Wildasin, D., April 1985. An alternative approach to aggregate surplus analysis. Journal of Public Economics 26 (3), 289–302.

Hartwick, J., February 1978. Optimal price discrimination. Journal of Public Economics 9 (1), 83–89.

Hatta, T., February 1977. A theory of piecemeal policy recommendations. Review of Economic Studies 44 (1), 1–21.

Mirrlees, J., April 1971. An exploration in the theory of optimal income taxation. Review of Economic Studies 38 (2), 175–208.

Mirrlees, J., February 1975. Optimal commodity taxation in a two-class economy. Journal of Public Economics 4 (1), 27–33.

Mirrless, J., November 1976. Optimal tax theory: a synthesis. Journal of Public Economics 6 (4), 327–358.

Sadka, E., June 1976. On income distribution, incentive effects, and optimal income taxation. Review of Economic Studies 43 (2), 261–268.

Samuelson, P.A., February 1956. Social indifference curves. Quarterly Journal of Economics 70 (1), 1–22.

Seade, J., April 1977. On the shape of optimal tax schedules. Journal of Public Economics 7 (2), 203–235.

Sheshinski, E., July 1972. The optimal linear income tax. Review of Economic Studies 39 (3), 297–302.

Stern, N., July/August 1976. On the specification of models of optimum income taxation. Journal of Public Economics 6 (1-2), 123–162.

Stiglitz, J., Dasgupta, P., April 1971. Differential taxation, public goods and economic efficiency. Review of Economic Studies 38 (2), 151–174.

Chapter 15

Taxation under Asymmetric Information

Chapter Outline

The text so far has ignored an important market imperfection, the presence of private or asymmetric information. Chapter 15 explores the implications of asymmetric information on taxation, and later chapters extend the analysis to transfer payments and other public expenditures. Analysis under the assumption of asymmetric information is inherently second best because first-best analysis requires that agents have perfect information about everything relevant to their economic decisions and exchanges.

The problem of asymmetric information has been a focal point of public sector analysis for the past 30–40 years, just as it has been in almost all fields of economics. The recent interest in asymmetric information is understandable, first and foremost because it is so common. Agents often possess private information about themselves that other agents, including the government, do not or cannot know, at least not without undertaking considerable effort and cost to monitor behavior. In addition, optimizing agents have obvious incentives to exploit private information to their own advantage, and economists quite naturally assume they will do so to the fullest extent possible. Finally, the assumption of asymmetric information often produces results that are very different from those obtained under the assumption of perfect information. This has been especially true in public sector economics.

Regarding the theory of taxation, old standards such as the Ramsey tax rule for one-consumer equivalent economies or the Diamond–Mirrlees many-person tax rule are no longer prescriptions for optimal taxation under private information. This is so even if the government can in principle tax (almost) everything as the Ramsey and Diamond–Mirrlees models assume. In fact, governments may be quite restricted in what they can tax under private information. They may not have sufficient information about some economic variables to use them as a tax base. What the government can tax is an important policy question in a world of private information. A final point is that private information can severely constrain a government's ability to redistribute purchasing power through taxes and transfers.

The last comment on redistribution points to a special difficulty with private information: It is not simply a

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00015-7

technological or structural imperfection. Rather, it has certain uncomfortable behavioral characteristics that are absent from other market imperfections such as distorting taxes, monopoly power, or legislated budget constraints. Agents who exploit private information for their own self-interest at the expense of broader social goals tear at the fabric of society. They violate the spirit of good citizenship that is necessary to hold a society together. An obvious example is people who evade paying taxes that would have been transferred to the poor.[1]

Such behavior also strikes at the foundations of normative public sector theory. What is normative theory to make of the tax evaders, especially when the norms include a concern for equity as well as efficiency? In formal terms, should the social welfare function give dishonest taxpayers the same ethical weight as honest taxpayers who could also exploit private information but do not? If unequal weights are chosen, how unequal should they be? Should the dishonest receive zero weight? If the dishonest are to receive a positive weight, should society be expected to spend its scarce resources on monitoring their behavior or on punishing them? Should the government significantly alter its tax policies to discourage dishonest behavior? These are difficult, open questions that have profound implications for any normative analysis. Different answers can lead to very different policy recommendations.

As it happens, much of the recent normative public sector analysis has continued the long-established tradition of using the equal-weight utilitarian social welfare function as the objective function, in essence simply adding the assumption of asymmetric information to existing models. This strategy makes sense for studying the other-things-equal implications of asymmetric information. At the same time, however, it implicitly condones the incentive to exploit private information as natural and socially acceptable behavior. Is this sensible as a basis for normative policy analysis? If not, what should the social objective function be? The foundations of normative theory, always vulnerable at best, become shaky indeed in the face of private information.

Private information can greatly complicate the quest for end-results equity, so let us begin with the distribution question.

LUMP-SUM REDISTRIBUTIONS AND PRIVATE INFORMATION

In a first-best environment, with perfect information, the government should transfer one good or factor lump sum to satisfy the interpersonal equity conditions for a social welfare maximum. Presumably high-ability, high-income people would be taxed and low-ability, low-income people would receive transfers.

Lump-sum redistributions on the basis of ability could raise serious objections, however. Suppose, as is commonly assumed, that everyone has the same tastes and the social welfare function is equal weighted in the sense that everyone has the same social marginal utility at the same commodity bundle. People differ only in their abilities. Under these assumptions, the optimal lump-sum redistributions may violate Feldstein's vertical equity principle of no reversals. The high-ability individuals, who are clearly better off before the redistributions, may be worse off at the social welfare optimum after the redistributions. The following simple example illustrates the possibility of reversals.[2]

Suppose there are two types of people: high-ability people (H) who receive a wage W_H and low-ability people (L) who receive a wage W_L, $W_H > W_L$. The two types of people have identical utility functions defined over a composite commodity, C, and labor, L, with L measured negatively. Assume further that utility is separable in consumption and labor:

$$U(C, L) = f(C) + g(L) \tag{15.1}$$

with L measured negatively. Markets are competitive, consistent with a first-best environment, and $P_C = 1$, the numeraire.

- Pareto optimality: The two types of people equate their marginal rates of substitution between consumption and leisure to their wages, as required for pareto optimality:

$$\frac{g_{L_H}}{f_{C_H}} = W_H; \quad \frac{g_{L_L}}{f_{C_L}} = W_L \tag{15.2}$$

- Interpersonal equity: Assuming that the good C is redistributed lump sum, the redistribution equalizes the social marginal utility of consumption across the two types of people (and within each type):

$$\frac{\partial W}{\partial U^H} f_{C_H} = \frac{\partial W}{\partial U^L} f_{C_L} \tag{15.3}$$

With an equal-weighted social welfare function and the same tastes, Eqn (15.3) implies and $f_{CH} = f_{CL}$ and $C_H = C_L$ at the social welfare optimum. But equal consumption, coupled with the pareto-optimal condition, Eqn (15.2), implies $g_{L_H} > g_{L_L}$. The marginal disutility of work is greater for the high-ability types; they work harder. Therefore, the high-ability people have the same level of

1. The same could be said of agents who exploit monopoly power, but their behavior is not different in kind from profit maximizing under perfect competition, unlike the distinction between honest and dishonest taxpayers. It is also less secretive than something like tax evasion.

2. The example is taken from Stiglitz (1987).

consumption as the low-ability people and work harder at the social welfare optimum. They are worse off after the lump-sum redistributions, in violation of Feldstein's no-reversals principle.

The reversal solution is guaranteed in this example because of the separability assumption. It may not happen with more general, nonseparable utility, but it could, as illustrated in Fig. 15.1. A and A' are the equilibria for each high-ability person before and after the lump-sum tax, and B and B' are the corresponding equilibria for each low-ability person before and after the lump-sum transfer. A reversal is more likely the larger the redistributions required to satisfy the interpersonal equity conditions.

Almost everyone would object to a tax-transfer policy that leads to utility reversals. Therefore, although taxes and transfers based on ability are lump sum and first-best optimal, high-ability people have a strong incentive to hide their ability from the government. Assume they can do so. Given this incentive, a natural modeling strategy is to assume that the government can at best know people's incomes but not the separate components of their incomes, their wages or their hours worked. The wage is an index of ability, and knowing the hours worked, given income, would reveal the wage. But income is endogenous, so that taxes and transfers of income cannot be lump sum. Thus, the incentive and the means to hide ability force the government into a second-best trade-off between equity and efficiency, in which the equity gains of redistributing income must be balanced against the efficiency losses of both taxing and transferring the income. The redistributions of income must also guard against the possibility of reversals; if not, high-income, high-ability people have an incentive to represent themselves as low-ability people who have high incomes because they work extra hard.

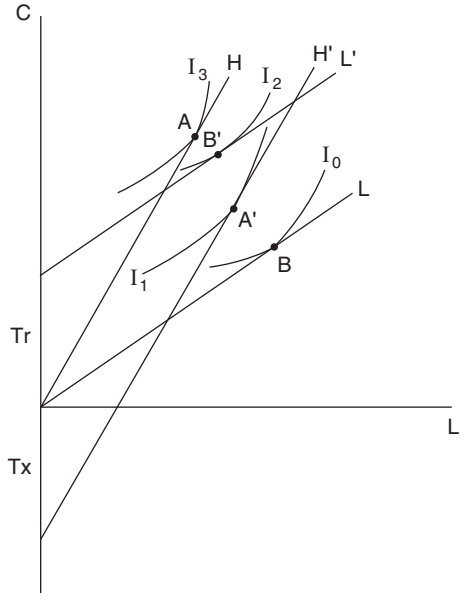

FIGURE 15.1

In summary, private information forces the government to rely on more restricted forms of taxation than are required to achieve the first-best interpersonal equity conditions. Bator's first-best bliss point is unattainable under private information, in general.

REDISTRIBUTION THROUGH COMMODITY TAXATIONS

One way to reduce the probability of reversals, and the resulting incentive to hide information from the government, is to rely solely on commodity taxes and subsidies.[3] Excise and sales taxes (subsidies) are levied on firms, not individuals, and firms' revenues may be easier to monitor than individual abilities or incomes. Collecting taxes from business is really the only choice when countries are in the early stages of economic development and literacy rates are low. A broad-based income tax requires that the population can keep the records and file the forms associated with an income tax. In contrast, the highly developed industrialized nations can easily use income taxes if they wish. Literacy rates exceed 90%, a very high percentage of economic activity is marketed, and firms can help administer the tax through withholding of income tax liabilities as the income is earned. The interesting question, however, is whether an industrialized nation should prefer commodity taxation to income taxation in a world of imperfect information.

A potential drawback to commodity taxation and subsidy is that it might not have much redistributional bite. The only way to redistribute purchasing power under a pure commodity tax/subsidy scheme is to tax the goods and services favored relatively more by high-income people and subsidize the goods and services favored relatively more by low-income people. This is clearly not as redistributive as directly taxing and transferring incomes unless the rich and poor buy vastly different goods and services.

Raaj Sah developed a simple and ingenious method for determining the limits of redistribution under commodity taxation that relies only on the government's budget constraint (Sah, 1983). His method led him to conclude that commodity taxes and subsidies are unlikely to have much equalizing effect on the distribution of income.

Sah employs a standard many-person commodity tax model with linear technology (fixed producer prices). Labor, the only factor of production, is in fixed supply and is the untaxed numeraire. The taxes and subsidies are levied per unit, such that $q_i = P_i + t_i$, where q_i are the consumer prices and P_i are the producer prices. The government

3. In this chapter "commodities" has the standard meaning as goods and services.

budget constraint is $\sum_{i=1}^{N} t_i X_i = 0$, with X_i the aggregate quantity of good i, $t_i > 0$ for the taxed goods, and $t_i < 0$ for the subsidized goods. Setting total taxes equal to total subsidies is a convenience that highlights the distributional impact of the taxes and subsidies. Finally, Sah assumes a Rawlsian social welfare function, which has two advantages. As the most egalitarian of the social welfare functions, it generates the greatest possible incentive to redistribute. It also provides an unambiguous measure of the improvement in the distribution because all that matters is how much the real income of the worst-off individual has increased.

Sah chooses the Hicks' equivalent variation (HEV) as the measure of real income improvement, defined as

$$\text{HEV}^1 = M^1\left(p, V\left(q, I^1\right)\right) - I^1 \tag{15.4}$$

$M^1(\,)$ is the expenditure function, V is the indirect utility function, and I^1 is the fixed labor income of the worst-off individual, person 1.[4] The HEV is the lump-sum income the worst-off individual would be willing to pay to return to the pretax prices, p. The metric of distributional improvement is the proportional increase in the real income of the worst-off person:

$$\text{HEV}^1/I^1 = M^1\left(p, V\left(q, I^1\right)\right)/I^1 - 1 \tag{15.5}$$

The limits to the distributional improvement rely on the property that the expenditure function is quasi-concave in prices. Thus,

$$M^1\left(q, V^1\left(q, I^1\right)\right) + \nabla_q M^1\left(q, V^1\left(q, I^1\right)\right)(p - q) \geq M^1\left(p, V^1\left(q, I^1\right)\right) \tag{15.6}$$

See Fig. 15.2. Starting from q, a movement along the slope of M to p leaves the consumer above the value of the expenditure function at p. The first term on the left-hand side (LHS) of Eqn (15.6) is the value of the expenditure function at the actual with-tax equilibrium, equal to the worst-off individual's fixed income, I^1. The second term is $-Xt$, the net subsidy received by the worst-off individual, $-T^1$. Therefore,

$$I^1 - T^1 \geq M\left(p, V\left(q, I^1\right)\right) \tag{15.7}$$

Dividing by I^1 and using Eqn (15.5) yields

$$-T^1/I^1 \geq \text{HEV}^1/I^1 \tag{15.8}$$

Equation (15.8) says that the proportional improvement in the real income of the worst-off individual must be less than or equal to the ratio of his or her net subsidy to income.

Sah establishes limits on the ratio of net subsidy to income in terms of the overall government budget

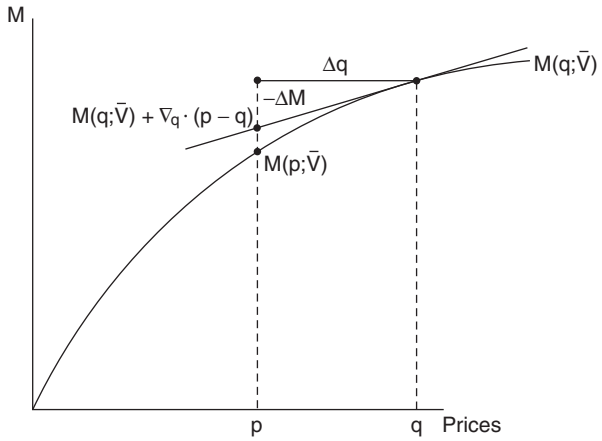

FIGURE 15.2

constraint, $\sum_{i=1}^{N} t_i X_i = 0$, written in terms of budget shares as follows.

Define $\phi_i = t_i/q_i$ as the proportional tax (subsidy) rate on good i in terms of the gross-of-tax price. Note for future reference that $\phi_i < 1$, since $t_i = q_i - P_i$. Also, define $I = \sum_{h=1}^{H} I^h$ as the total fixed labor income in the economy.

Multiplying and dividing each term in the government budget constraint by q_i and dividing the entire budget constraint by I yields

$$\sum_{i=1}^{N} \frac{t_i X_i}{I} = \sum_{i=1}^{N} \frac{t_i q_i X_i}{q_i I} = \sum_{i=1}^{N} \frac{\phi_i q_i X_i}{I} = \sum_{i=1}^{N} \phi_i W_i \tag{15.9}$$

where $W_i =$ the aggregate budget share for good i.

Next divide the goods into the subsets of taxed goods T, with $\phi_i > 0$, and subsidized goods S, with $\phi_i < 0$:

$$\sum_{ieT} \phi_i W_i = -\sum_{ieS} \phi_i W_i \tag{15.10}$$

But,

$$1 > \sum_{ieT} W_i > \sum_{ieT} \phi_i W_i \tag{15.11}$$

since $\phi_i < 1$. Therefore,

$$1 > -\sum_{ieS} \phi_i W_i \tag{15.12}$$

Next consider the $\min_j\{W_j\}$ such that $W_i/\min_j\{W_j\} \geq 1$, for all i. Then,

$$-\sum_{ieS} \phi_i W_i/\min_j\{W_j\} \geq -\sum_{ieS} \phi_i \tag{15.13}$$

Therefore, from Eqn (15.12),

$$-1/\min_j\{W_j\} > -\sum_{ieS} \phi_i \tag{15.14}$$

4. p and q are vectors of prices here.

Return to the worst-off individual, person 1:

$$-T^1 \Big/ I^1 = -\sum_{i=1}^{N} \phi_i W_i^1 = -\sum_{ieT} \phi_i W_i^1 - \sum_{ieS} \phi_i W_i^1 \tag{15.15}$$

where W_i^1 is person l's budget share of good i. Therefore,

$$-T^1 \Big/ I^1 < -\sum_{ieS} \phi_i W_i^1 \tag{15.16}$$

Also, $W_i^1 \leq \max_j \{W_j^1\}$. Therefore,

$$-T^1 \Big/ I^1 \leq -\sum_{ieS} \phi_i \max_j \left\{ W_j^1 \right\} \tag{15.17}$$

and from Eqn (15.14),

$$-T^1 \Big/ I^1 < \left[1 \Big/ \min_j \left\{ W_j \right\} \right] \max_j \left\{ W_j^1 \right\} \tag{15.18}$$

Comparing Eqn (15.18) with Eqn (15.8), the proportional improvement in the real income of the worst-off individual must be less than the ratio of his or her maximum budget share to the minimum economy-wide budget share. To push this limit as high as possible, assume that the richest person r has infinite income so that $W_i = W_i^r$, for all i. Then the limit depends on the maximum budget share of the worst-off individual and the minimum budget share of the richest individual. Suppose some necessity item is 80% of person l's budget and 20% of r's budget. Then the limit of the worst-off's gain in real income is four times his or her income. This may appear to be a large gain, but it is made under the extreme assumption of the richest person having infinite income. The minimum economy-wide budget share is likely to be much more than 1/4 as large as the maximum budget share of the worst-off individual. Also, this is the limit of gain for the worst-off individual, not the average poor person.

Sah conducts a number of simple exercises to get some feeling for the amount of redistribution that is likely through commodity taxes and subsidies. The variations include optimal taxation with a Rawlsian social welfare function; CES utility functions with varying degrees of elasticity of substitution, from Cobb−Douglas to Leontief; two classes of people with different preferences; uniform preferences with an arbitrary number of classes; wide differences in the range of incomes from richest to poorest; and one experiment using actual data for the United Kingdom and the linear expenditure system. The exercises almost always produce very modest proportional gains in the real income of the worst-off individual(s), usually less than 1.5 and often much less. Sah concludes that not much redistribution is likely to be possible through commodity taxes and subsidies.

The only caveat is if the rich consume some goods that the poor do not consume and vice versa. Then indirect taxes and subsidies can be targeted to the rich and poor just as income taxes can, with much greater redistributional impact. The only natural limitation on the amount of redistribution is the size of the tax base on the items consumed exclusively by the rich. For instance, how much revenue can the government raise from a tax on yachts?[5]

OPTIMAL TAXATION, PRIVATE INFORMATION, AND SELF-SELECTION CONSTRAINTS

Suppose society decides that it has to resort to direct taxes on income to achieve the redistributional bite that it wants from its tax system. It then has to confront the two problems with income taxes mentioned above. One is the trade-off between the distributional gains and the efficiency losses of taxing endogenous income. The other is the potential of violating Feldstein's vertical equity principle of no reversals. This section focuses on the second problem since it is a fundamental problem for income taxation in a world of private information no matter what norms the government is trying to pursue.

The principle of no reversals is so deeply held that the government might want to design its tax system to prevent reversals from occurring. Taxpayers have an incentive to hide income from the tax authorities under the best of circumstances. The incentive becomes especially strong if taxpayers fear that their ranking in the income distribution would be lower after taxes. An income tax might not even be viable if the potential for reversals is widespread.

No-reversal constraints take the form that each taxpayer prefers his after-tax bundle of goods and factors to anyone else's after-tax bundle. As such, they are called self-selection constraints, because they ensure that taxpayers will reveal who they are to the tax authorities. Taxpayers may still try to hide income, but at least they will not claim to be someone else to reduce their tax liability. Self-selection constraints are also called incentive compatibility constraints because the utility maximizing strategy under the constraints is for taxpayers to reveal their true

5. The analysis in this section should not leave the impression that commodity taxes necessarily avoid problems of imperfect information. The diversion of goods to black markets in an effort to escape taxation is always a potential problem, especially in low-income countries. John McLaren has published an analysis of optimal commodity taxes when evasion through black markets is possible. His model generates a number of interesting conclusions. One of the more compelling is that the government might want to tax just one good rather than use a broader based sales tax to save on enforcement costs. This is exactly what many of the poorer developing economies choose to do when they begin to levy taxes. Later on in this chapter we present an analysis of tax evasion under an income tax. Space limitations prevent a presentation of McLaren's model as well, but interested readers should consult McLaren (1998).

identities. That is, the incentive to tell the truth is compatible with utility maximization. Incentive compatibility is a fundamental goal of the theory of mechanism design in the presence of imperfect information.

Once the government designs self-selection constraints into the tax system, the problem of reversals is no longer just a matter of equity. The constraints become part of whatever second-best problem the government is trying to solve, whether it be maximizing social welfare or a more restricted goal such as second-best pareto efficiency or revenue maximization. We will consider the problem of achieving second-best pareto efficiency under self-selection constraints to highlight the effects of the constraints on the design of optimal taxes.

Elements of the Model

The no-reversal, self-selection constraints happen to have a profound effect on the design of taxes. This can be seen by modifying one of the many-person models in Chapter 14 to include income taxation and the self-selection constraints. The simplest choice is a model with N commodities (goods and services), labor as the only factor of production, and linear technology with fixed producer prices. For convenience, the quantities of each commodity are defined such that all producer prices equal 1. Defining producer prices for each commodity would add no insights, and the notational demands of the model are heavy enough as is. As in the earlier discussion of lump-sum taxation, there are two classes of people, those with high ability (H) and those with low ability (L), who receive wages W_H and W_L, respectively. Everyone has identical preferences; people differ only in their abilities.

Regarding taxation, the tax authorities can monitor income perfectly but not ability, that is, not the wages or the hours worked. Therefore, they cannot know who has high ability and who has low ability. Taxes can be levied on any of the commodities and on income (but not labor or wages separately). Moreover, all taxes can potentially be nonlinear to enhance their distributional power.[6] Finally, the model incorporates the self-selection constraints to ensure that the two classes of people reveal themselves to the tax authorities.

Preferences

Given that income is taxed, define preferences in terms of income rather than labor:

$$V^h = V^h(X_{hj}; Y^h) \quad h = H, L; j = 1, \ldots, N$$

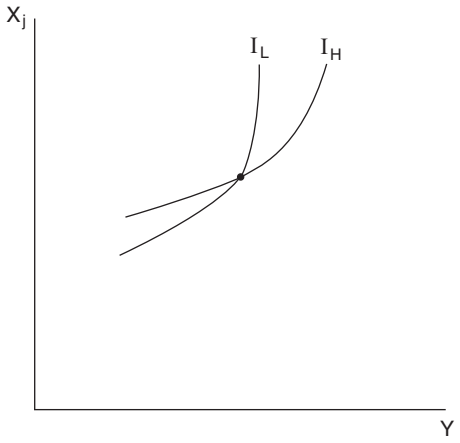

FIGURE 15.3

where

X_{hj} = commodity j purchased by a person of ability h

Y^h = the income of a person of ability h, with $Y^h = W_h L_h$.

Four properties of V are worth noting:

1. Income is a bad, not a good, in this specification because each person has to supply more labor at the fixed wages to earn more income. Therefore, the indifference curves for one of the commodities X_{hj} and Y^h are upward sloping as they would be if labor were on the horizontal axis.
2. The assumption that everyone has the same tastes implies that the indifference curves in (X_j, L) space are the same for both classes of people. But the indifference curves in (X_j, Y) space differ for the two classes. They are flatter for the high-ability people at a given (X_j, Y), as pictured in Fig. 15.3. This follows because

$$\begin{aligned} V^h &= V^h(X_{hj}; Y^h) = U^h(X_{hj}; W_h L_h / W_h) \\ &= U^h(X_{hj}; Y^h / W_h) \end{aligned}$$

Therefore, the marginal rate of substitution in terms of one of the commodities X_j and Y is

$$-(dX_{hj}/dY^h)_{V=\overline{V}} = -\frac{1}{W_h}(dX_{hj}/dL_h)_{U=\overline{U}} \qquad (15.19)$$

Consumers require only $1/W_h$ as much additional X_{hj} to be indifferent to a unit increase in Y^h as they would to a unit increase in L_h. Also, since $W_H > W_L$, the marginal rate of substitution (MRS) is flatter for the high-ability people at the same (X_J, Y) point. Intuitively, the high-ability people are willing to accept less additional X_j to compensate for an additional unit of Y because they have to supply less labor (can enjoy more leisure) to obtain the same amount of Y.

6. A model of this nature with self-selection constraints was first developed by A. Atkinson and J. Stiglitz in 1976. See Atkinson and Stiglitz (1976). The analysis here closely follows the presentation by Stiglitz in the *Handbook of Public Economics*. See Stiglitz (1987), pp. 991–1041.

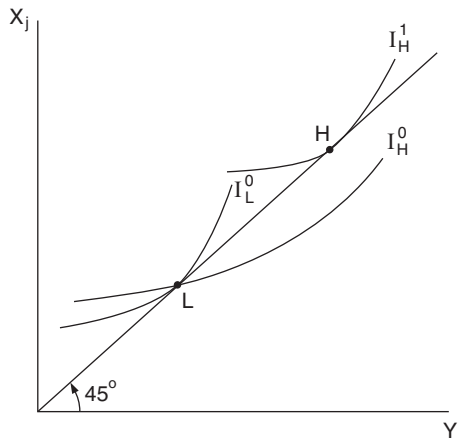

FIGURE 15.4

3. The zero-tax consumer equilibrium requires that the MRS between any commodity and labor be equal to the wage (with all commodity prices equal to one):

$$-\left(\mathrm{d}X_{hj}/\mathrm{d}L_h\right)_{U=\overline{U}} = W_h \tag{15.20}$$

Therefore, the zero-tax equilibrium condition in terms of any commodity and income is

$$-\left(\mathrm{d}X_{hj}/\mathrm{d}Y^h\right)_{V=\overline{V}} = 1 \tag{15.21}$$

A possible zero-tax equilibrium is pictured in Fig. 15.4. Notice that the zero-tax competitive equilibrium is incentive compatible, as would be a first-best equilibrium with lump-sum (nondistorting) taxes.

Self-Selection Constraints

The self-selection constraints require that each class prefers its own bundle of commodities and income:

$$V^H(X_{HJ}; Y^H) \geq V^H(X_{LJ}; Y^L) \tag{15.22}$$

and

$$V^L(X_{LJ}; Y^L) \geq V^L(X_{HJ}; Y^H) \tag{15.23}$$

The realistic concern is that the high-ability class will prefer the low-ability bundle to avoid tax liability. The low-ability class is unlikely to pretend to be of high ability because they would have to sacrifice too much leisure to earn Y^H at a wage of W_L. Therefore, only the high-ability constraint is likely to bind in a tax equilibrium.

Government Budget Constraint

Suppose the government has to raise a fixed amount of revenue, R. If there are N^H high-ability people and N^L low-ability people, then the government's budget constraint is

$$N^H\left(Y^H - \sum_{j=1}^{N} X_{Hj}\right) + N^L\left(Y^L - \sum_{j=1}^{N} X_{Lj}\right) = R \tag{15.24}$$

in terms of the X_{hj} and Y^h. The government has access to all income that is not consumed.[7]

Pareto-Efficient Taxation

The goal is to determine the pareto-efficient pattern of commodity and income taxes that raises a given amount of revenue, such that the self-selection constraints hold. The search is for a utility-possibilities frontier rendered second best by the private information on abilities that forces the government to impose the self-selection constraints. Since everyone within each ability class is identical, the frontier is the set of allocations that maximizes the utility of a representative person in one of the classes subject to holding the utility of a representative person in the other class constant and subject to the self-selection and revenue constraints. Formally, the problem is to

$$\max_{\left\{X_{Hj};X_{Lj};Y^H;Y^L\right\}} V^H\left(X_{Hj}; Y^H\right)$$

$$\begin{aligned}
\text{s.t}\quad & V^L\left(X_{Lj}; Y^L\right) = \overline{V}^L \\
& V^H\left(X_{Hj}; Y^H\right) \geq V^H\left(X_{Lj}; Y^L\right) \\
& V^L\left(X_{Lj}; Y^L\right) \geq V^L\left(X_{Hj}; Y^H\right) \\
& N^H\left(Y^H - \sum_{j=1}^{N} X_{Hj}\right) + N^L\left(Y^L - \sum_{j=1}^{N} X_{Lj}\right) = R
\end{aligned}$$

Defining multipliers for the constraints, the Lagrangian is

$$\begin{aligned}
\max_{\left\{X_{Hj};X_{Lj};Y^H;Y^L\right\}} L = \; & V^H\left(X_{Hj}; Y^H\right) + \mu\left(V^L\left(X_{Lj};\ Y^L\right) - \overline{V}^L\right) \\
& + \lambda^H\left(V^H\left(X_{Hj}; Y^H\right) - V^H\left(X_{Lj}; Y^L\right)\right) \\
& + \lambda^L\left(V^L\left(X_{Lj}; Y^L\right) - V^L\left(X_{Hj}; Y^H\right)\right) \\
& + \gamma\left(N^H\left(Y^H - \sum_{j=1}^{N} X_{Hj}\right)\right) + \left(N^L\left(Y^L - \sum_{j=1}^{N} X_{Lj}\right) - R\right)
\end{aligned}$$

The first-order conditions are

$$\begin{gathered}
X_{Hj}\colon \partial V^H/\partial X_{Hj} + \lambda^H \partial V^H/\partial X_{Hj} - \lambda^L \partial V^L/\partial X_{Hj} - \gamma N^H = 0 \\
j = 1, \ldots, N
\end{gathered} \tag{15.25}$$

7. Dollars of pretax income and the quantities of the commodities are in the same dollar units given the pricing convention.

$$X_{Lj}\colon \mu\partial V^L/\partial X_{Lj} - \lambda^H\partial V^H/\partial X_{Lj} + \lambda^L\partial V^L/\partial X_{Lj} - \gamma N^L = 0 \quad j = 1, \ldots, N \tag{15.26}$$

$$Y^H\colon \partial V^H/\partial Y^H + \lambda^H\partial V^H/\partial Y^H - \lambda^L\partial V^L/\partial Y^H + \gamma N^H = 0 \tag{15.27}$$

$$Y^L\colon \mu\partial V^L/\partial Y^L - \lambda^H\partial V^H/\partial Y^L + \lambda^L\partial V^L/\partial Y^L + \gamma N^L = 0 \tag{15.28}$$

Rearrange terms and divide pairs of the first-order conditions in the usual manner to derive the four relevant sets of MRSs:

$$\mathrm{MRS}^H_{XHj,XHk} = \frac{\partial V^H/\partial X_{HK}}{\partial V^H/\partial X_{Hj}} = \frac{-\lambda^H\partial V^H/\partial X_{HK} + \lambda^L\partial V^L/\partial X_{HK} + \gamma N^H}{-\lambda^H dV^H/\partial X_{Hj} + \lambda^L\partial V^L/\partial X_{Hj} + \gamma N^H} \quad \text{all } j,k \tag{15.29}$$

$$\mathrm{MRS}^L_{XLj,XLk} = \frac{\partial V^L/\partial X_{LK}}{\partial V^L/\partial X_{Lj}} = \frac{+\lambda^H\partial V^H/\partial X_{LK} - \lambda^L\partial V^L/\partial X_{LK} + \gamma N^L}{+\lambda^H\partial V^H/\partial X_{Lj} - \lambda^L\partial V^L/\partial X_{Lj} + \gamma N^L} \quad \text{all } j,k \tag{15.30}$$

$$\mathrm{MRS}^H_{XHj},\, Y^H = \frac{\partial V^H/\partial Y^H}{\partial V^H/\partial X_{Hj}} = \frac{-\lambda^H\partial V^H/\partial Y^H + \lambda^L\partial V^L/\partial Y^H - \gamma N^H}{-\lambda^H\partial V^H/\partial X_{Hj} + \lambda^L\partial V^L/\partial X_{Hj} + \gamma N^H} \quad \text{all } j,k \tag{15.31}$$

$$\mathrm{MRS}^L_{XLj},\, Y^L = \frac{\partial V^L/\partial Y^L}{\partial V^L/\partial X_{Lj}} = \frac{+\lambda^H\partial V^H/\partial Y^L - \lambda^L\partial V^L/\partial Y^L - \gamma N^L}{+\lambda^H\partial V^H/\partial X_{Lj} - \lambda^L\partial V^L/\partial X_{Lj} + \gamma N^L} \quad \text{all } j,k \tag{15.32}$$

Consider the following two cases.

Self-Selection Constraints Not Binding

Suppose that neither self-section constraint is binding so that $\lambda^H = \lambda^L = 0$. Then all the relevant MRS = 1.[8] There should be no distorting taxation, the first-best result, because the private information does not truly constrain the government. This does not necessarily rule out taxation of the commodities and income if taxes are nonlinear, only that the marginal tax rates must be zero. Average rates of tax could be positive, in which case they would be equivalent to lump-sum taxes because they do not affect decisions on the margin.

Self-Selection Constraint on the High-Ability Class Binding

As noted above, the realistic case is for the self-selection constraints to bind on the high-ability class but not the low-ability class so that $\lambda^H > 0$ and $\lambda^L = 0$.

High-Ability Class

Consider, first, the MRSs for the high-ability class. A remarkable result is immediately evident. $\mathrm{MRS}^H X_{Hj}$, $X_{Hk} = \mathrm{MRS}^H X_{Hj}$, $Y^H = 1$. All marginal tax rates on the high-ability class should be zero; they should face no distorting commodity or income taxation. Again, this does not rule out taxation of the rich if taxes are nonlinear, just nonzero marginal taxes. The average rates of tax could be positive. This result is quite robust, applying to models with many classes of taxpayers and even a continuum of taxpayers. The marginal tax rates on the highest ability, highest income taxpayers should be zero.

The intuition for the zero marginal income tax rate is as follows. We have seen that the high-ability taxpayer would set $\mathrm{MRS}^H X_{Hj}$, $Y^H = 1$. With a positive marginal tax rate, T′, the taxpayer would set $\mathrm{MRS}^H X_{Hj}$, $y^H = (1 - \mathrm{T}') < 1$. Suppose the marginal tax rate is very small, so that by setting it equal to zero the taxpayer essentially moves along the indifference curve. Since the slope is less than one, income rises more than consumption, which generates some tax revenue to give to the low-ability taxpayer. Therefore, the move from a positive to a zero marginal tax rate leaves the high-ability taxpayer indifferent while increasing the utility of the low-ability taxpayer. The positive tax rate cannot be pareto efficient. (The same argument applies in reverse for a small marginal subsidy. Removing it is pareto improving.)

The zero-marginal-tax-rate result stands in direct contrast to the result obtained in the many-person linear commodity tax model in Chapter 14. In that model, commodities consumed relatively more by the high-income classes (lower social marginal utilities of income) are taxed at higher (marginal) rates.[9] The result also begins to suggest some of the difficulties that private information causes for public policy. Although people undoubtedly want to avoid after-tax reversals, they are unlikely to embrace a tax system with zero marginal tax rates on the richest citizens.

Low-Ability Class

The same results do not apply to the low-ability taxpayers.

Income Taxation. Consider first income taxation, which is based on the MRS between one of the goods and income. With $\lambda^H > 0$ and $\lambda^L = 0$, Eqn (15.32) becomes

$$\frac{\partial V^L/\partial Y^L}{\partial V^L/\partial X_{Lj}} = \frac{+\lambda^H\partial V^H/\partial Y^L - \gamma N^L}{+\lambda^H\partial V^H/\partial X_{Lj} + \gamma N^L} \quad \text{all } j \tag{15.33}$$

8. In absolute value. The MRS between two goods and between any one good and income have opposite signs, since income is a "bad."

9. Recall that the taxes considered in Chapter 14 were linear.

Suppose the government levies a general income tax $T = T(Y)$, with marginal tax rate T'. The low-ability taxpayer would set

$$(-)\frac{\partial V^L/\partial Y^L}{\partial V^L/\partial X_{Lj}} = (1 - T') \tag{15.34}$$

The question, then, is what does the right-hand side (RHS) of Eqn (15.33) imply about T'? To sign the RHS of Eqn (15.33), define:

$$a^h = -\left(\partial V^h/\partial Y^L\right)\Big/\left(\partial V^h/\partial X_{Lj}\right) = (\mathrm{d}X_{Li}/\mathrm{d}Y_L)_{V^h=\overline{V}^H}$$
$$h = H, L$$

and

$$v = \lambda^H\left(\partial V^H/\partial X_{Lj}\right)/\gamma N^L$$

Note that $v > 0$ since every term in v is positive. Divide the numerator and denominator of the RHS of Eqn (15.33) by γN^L and note the sign change in the numerator to obtain

$$(-)\frac{\partial V^L/\partial Y^L}{\partial V^L/\partial X_{Lj}} = \frac{\lambda^H\left(\partial V^H/\partial Y^L\right)\gamma N^L + 1}{+\lambda^H\left(\partial V^H/\partial X_{Lj}\right)\gamma N^L + 1} \tag{15.35}$$

Substituting for a^h and v, Eqn (15.35) becomes

$$a^L = (1 + va^H)/(1 + v) \tag{15.36}$$

Adding and subtracting a^H in the numerator and simplifying yields

$$a^L = a^H + (1 - a^H)/(1 + v) \tag{15.37}$$

But $a^L > a^H$ since the MRS is steeper for the low-ability individual at the same (Y^L, X_{Lj}). Therefore, $(1 - a^H)/(1 + v) > 0$, so that $a^H < 1$. But $a^H < 1$ implies $a^L < 1$ from Eqn (15.36), and $a^L = (1 - T')$. Hence, $T' > 0$. The marginal income tax rate on the low-ability class should be positive.

Commodity Taxation. Consider, finally, the MRSs between the commodities:

$$\frac{\partial V^L/\partial X_{LK}}{\partial V^L/\partial X_{Lj}} = \frac{\lambda^H\partial V^H/\partial X_{LK} + \gamma N^L}{\lambda^H\partial V^H/\partial X_{Lj} + \gamma N^L} \tag{15.38}$$

The marginal tax rates are nonzero, in general. One notable exception is the case in which preferences are weakly separable between labor and the commodities, such that $\partial^2 V^h/\partial X_{hk}\partial L_h = 0$, all k, and $h = H, L$. Since everyone has identical preferences, weak separability implies that $\partial V^L/\partial X_{Lk} = \partial V^H/\partial X_{Lk}$ and $\partial V^L/\partial X_{Lj} = \partial V^H/\partial X_{Lj}$ at any given X_{Lk} and X_{Lj}. Substituting these equations into Eqn (15.38) implies

$$\frac{\partial V^L/\partial X_{LK}}{\partial V^L/\partial X_{Lj}} = 1 \tag{15.39}$$

The low-ability class should not face distorting commodity taxes, nor should the high-ability class (whether preferences are separable or not). Therefore, weakly separable utility in labor is a sufficient condition for levying only distorting income taxes, and then only on the low-ability class.

In the general case of nonseparable utility, the first-order conditions can be manipulated to show that the relative taxation of commodity j to commodity k depends upon the relative values of the MRSs between j and k for the high- and low-ability classes. The higher the relative MRS for the high-ability class, the higher the relative tax on j.[10] The intuition turns on the nature of the self-selection constraint. Only the low-ability class faces distorting taxes. Nonetheless, taxing more heavily the commodities favored relatively more by the high-ability class makes the low-ability class's commodity bundle less attractive to the high-ability class. This has the effect of relaxing the self-selection constraint and pushing out the second-best utility-possibilities frontier. Note how different this justification for higher or lower taxes is from the justification in the many-person Diamond–Mirrlees optimal commodity tax problem with its linear taxes and perfect information regarding people's ability. About the only point of similarity between the two models is their agreement that the pattern of commodity taxes depends importantly on the relationship between labor (leisure) and commodities in the consumers' preferences.

An Extension: The Direct–Indirect Tax Mix

Governments in the developed market economies typically choose a mix of indirect and direct taxes, for reasons that are not at all obvious. A common explanation is that a mix of taxes allows the governments to keep the rates low on each set of taxes, but consumers presumably understand that the combined weight of the indirect and direct taxes is what affects their welfare. The fact that each of the rates is low is more or less irrelevant. Furthermore, the Atkinson–Stiglitz model is not a helpful guideline for determining the optimal mix of indirect and direct taxes in the presence of private information. It has two main results to offer the policy maker. One is that only an income tax should be used if preferences are weakly separable between labor and the commodities. The other is that a mix of indirect and direct taxes should be used, but then the model only provides information on the marginal tax rates, not the average rates. Many different combinations of nonlinear commodity and income taxes could be used to meet the government's revenue needs.

Robin Boadway et al. developed a simple extension of the Atkinson–Stiglitz model that shows promise as a first step toward developing a theory of the optimal indirect–direct tax mix (Boadway et al., 1994). They note that a mix of taxes may be desirable because the income tax is easier to evade. They extend the two-class model to

10. The manipulations are tedious. They can be found in Stiglitz (1987), pp. 1025–1026.

include the possibility that the high-ability class can evade a portion of their income tax liability with no risk that the evasion can be detected. (The standard analysis of tax evasion assumes there is a chance of being caught; see later discussion.) Evaders do, however, bear a cost that depends on the proportion of the income evaded.

The model becomes extremely complicated with the addition of tax evasion, so much so that we will simply note three results of interest:

1. If commodity taxes are not used, then the possibility of evasion does not change the standard result that the marginal tax rate is zero on the high-ability people and positive on the low-ability people.
2. Starting from zero commodity taxes, the imposition of a uniform commodity tax is welfare improving.
3. Given an income tax, a uniform commodity tax is optimal if preferences are separable between the commodities and leisure and also quasihomothetic in the commodities. The model is too complex to lead to a simple characterization of the optimal pattern of commodity taxes when this condition is not satisfied and differentiated commodity taxation is called for.

In conclusion, the analysis in the section underscores the important point that the design of a tax system depends crucially on what the government can tax and what form the taxes can take. Both issues depend in large part on the information available to the government.

OPTIMAL INCOME TAXATION

The analysis of optimal income taxation was a natural research topic in the 1970s for economists interested in the properties of restricted taxation. The personal income tax had become the single most important tax in most of the developed market economies, as well as the main tax instrument for redistributing purchasing power. As such, the income tax was the obvious candidate for exploring the equity-efficiency trade-off in taxation. [11]

On the equity side, the income tax has considerable redistributive power because it can be so easily tailored to the personal and economic characteristics of individuals and families through features such as personal exemptions to protect the poor and graduated tax rates that tax high incomes in a very progressive manner. On the efficiency side, the income tax suffers from Okun's leaky bucket with its three main sources of leaks:

1. Dead-weight losses in labor and capital markets caused by the distorting nature of the tax.
2. Administrative costs of collecting the revenues, including the costs of monitoring taxpayers and enforcing the tax laws.
3. Compliance costs incurred by the taxpayers, both the costs of keeping records and filing the tax forms, whether by the taxpayer or a third-party tax preparer, and the costs incurred by taxpayers to reduce their tax liabilities, whether legally or illegally.

The optimal income tax is the one that achieves the optimal balance between the gains from redistribution as measured by some social welfare function and the three inefficiency costs from raising the tax revenue.

The first complete formal analysis of optimal income taxation was by James Mirrlees, and it stands as one of the classics in the public sector literature. It was the first formal model to incorporate the inefficiencies of the income tax in a social welfare framework. It was also the seminal article on the implications of private information on taxation. Mirrlees modeled only the labor market inefficiencies of the three leaks in Okun's bucket. He assumed that people varied by ability, but that the government could not know an individual's ability for the purposes of taxation. Instead, it was forced to tax income, not wages (ability) or labor supply separately. His model also implicitly honored the self-selection constraint because it explicitly incorporated the assumption that individuals would maximize their utility subject to the income tax function. Hence, they would necessarily prefer their own bundles of consumption and leisure to anyone else's bundle.

Mirrlees, and much of the literature that followed, specified a model with a continuum of taxpayers. These continuous models require the calculus of variations to solve, which is beyond the scope of this text. Nonetheless, the structure of the standard optimal income tax model pioneered by Mirrlees is easy enough to understand.

Stripped to its bare essentials, the optimal income tax problem can be represented as follows. Suppose each consumer has a preference function defined over the consumption of a composite commodity, c, and labor, 1:

$$U = U(c, 1) \tag{15.40}$$

11. The mainstream view of redistribution was (and still is) that it is a negative sum game because of the efficiency losses from the distorting taxes and transfers (i.e., Okun's leaky bucket). Some recent literature is trying to recover an older notion in political economy that redistribution can be a positive sum game because the transfers improve the productivity of the poor. This idea is particularly persuasive in some less-developed countries in which the poor may have such nutritionally poor diets absent redistribution that they do not have the strength to work. The possibility of positive-sum redistribution is also gaining support in the context of developed market economies where education is increasingly seen as the driver of productivity increases and hence long-run economic growth. Hoff and Lyon developed a model in which the distortions from wage taxation are more than offset by the productivity gains of the subsidies they finance, at least at sufficiently low levels of taxes and subsidies. The subsidies increase productivity because they are targeted to the education of the poor, who would otherwise be shut out of the market for higher education by market imperfections. See Hoff and Lyon (1995).

All individuals have identical preferences but varying abilities or skills indexed by the parameter n.[12] n transfers one unit of labor, 1, into nl efficiency units, which are assumed to be perfect substitutes in the production of c. Let w be the wage rate per efficiency unit of labor. Hence, an n-person's income is equal to

$$y = wnl \tag{15.41}$$

Assume further that the index of skills N is distributed across the population in accordance with a probability density function $\int_0^\infty (f(n)dn)$.

The government is interested in maximizing the Atkinson version of the continuous Bergson–Samuelson social welfare function of the form

$$W = \frac{1}{v}\int_0^\infty U^v(c,\ l)f(n)\mathrm{d}n \tag{15.42}$$

where v defines society's aversion to inequality, $v=1$ implies utilitarianism, and $v=-\infty$ implies the Rawls criterion of maximizing the utility of the individual with lowest utility.

The policy instrument is an income tax schedule of the general form

$$T = T(y) \tag{15.43}$$

with $T'>0$. $T(y)$ is assumed to be a general nonlinear schedule with, possibly, graduated marginal tax rates and subsidies to consumers below some threshold income level. In other words, the standard optimal income tax model is really a fully specified redistribution model of optimal income taxation and transfer. A common variation of $T(y)$ is the so-called credit income tax, a two-part schedule consisting of a fixed subsidy ("credit") and a constant marginal tax rate, $T=-\alpha+\beta y$. The government levies the income tax schedule to satisfy an aggregate budget constraint of the form

$$\int_0^\infty T(wnl)f(n)\mathrm{d}n = R \tag{15.44}$$

R could reflect some public goods or the deficits from decreasing cost production. $R=0$ implies that the government is solely interested in redistributing income.

Under the income tax, each individual has after-tax or transfer income available for consumption equal to

$$c = y - T(y) = wnl - T(wnl) \tag{15.45}$$

Each person maximizes utility Eqn (15.40) with respect to 1, given Eqn (15.45). The first-order condition is

$$wn(1-T')U_c + U_1 = 0 \tag{15.46}$$

so that the MRS between consumption and labor equals the after-tax (transfer) marginal wage. As noted earlier, Eqn (15.46) is the incentive compatibility constraint in the model.

The government's problem, then, is to maximize the social welfare function, Eqn (15.42), with respect to the parameters of $T(y)$, subject to the government budget constraint, Eqn (15.44), and the consumer equilibrium condition, Eqn (15.46). Equation (15.46) highlights the second-best nature of the problem, that the marginal tax rate T' distorts each consumer's choice between consumption and labor (leisure). With an Atkinson equal-weighted social welfare function and consumers having identical preferences, the first-best interpersonal equity conditions would imply equal posttax (transfer) income for all. If the income tax were lump sum, either because the labor supply was fixed or the government could tax and transfer on the basis of ability, the optimal marginal tax rate would be 100%. But, with variable labor supply and private information about ability, increases in marginal tax rates increase the distortion or efficiency loss, thereby partially offsetting the gains from an improved distribution. The optimal solution, then, finds the tax parameters that just equalize the efficiency losses and distributional gains on the margin.

Even the simplest optimal income tax model yields a number of interesting results. The components of the optimal tax schedule clearly depend on the structure of the tax schedule (e.g., linear or general) and the values of the parameters of the model, including the aversion to inequality, v; the distribution of skills throughout the population; the elasticity of labor supply; and the revenue requirement, R. Numerical analysis with a linear tax schedule has yielded a number of intuitively appealing results. Generally speaking, the marginal tax rate is higher:

1. The higher society's aversion to inequality (Atkinson, 1973; Stern, 1976): The more to be gained from redistribution, the more inefficiency society can tolerate.
2. The greater the dispersion of skills (Mirrlees, 1971; Stern, 1976): With an individualistic social welfare function, increased dispersion increases the gains from redistribution.
3. The lower the labor supply elasticity (Stern, 1976): Generally speaking, the efficiency loss implied by a given marginal tax rate varies directly with the labor supply elasticity. Nicholas Stern's simulation experiments showed that the marginal rate is extremely sensitive to the elasticity parameter, much more so than to any of the other parameters.

12. For our purposes, it does not matter whether these differing abilities are innate or the result of different educational experiences, so long as N is exogenous to each individual.

4. The higher the revenue requirement R (Stern, 1976): Roughly, a given tax rate entails less redistribution when some of the revenues must be used for other purposes. But this tends to increase the marginal returns from still further redistribution, implying a higher marginal rate.

The sensitivity of the optimal marginal tax rates to the labor supply elasticity deserves further comment. Mirrlees, and many of the other early income tax studies, assumed simple utility functions such as the Cobb−Douglas to get a feeling to the optimal tax rate. The utility functions chosen had very high labor supply elasticities (unity for Cobb−Douglas) that implied a relatively low marginal tax rate, on the order of 30%. This was much lower than the highest marginal tax rates in most of the developed market economies (70% in the United States at the time). Stern was the first to add a note of caution. He believed that the (compensated) labor supply elasticity was approximately 0.4, much lower than in the earlier models, which led him to propose a marginal tax rate of 54% for his most-preferred set of simulation parameters. The literature has never reached a consensus value for the labor supply elasticity, or therefore, for the dead-weight loss in the labor market for raising an additional dollar of income tax.

Subsequent studies of optimal income taxation have added the Okun leaks in the market for saving in a dynamic framework and in compliance costs in a static framework. No consensus has emerged on the marginal dead-weight loss from taxing income from saving, for the same reason as for labor. The estimates of the intertemporal elasticity of consumption are all over the place, from near zero to as high as 3. There is an emerging consensus on the combined administrative and compliance costs, which Joel Slemrod reports to be in the 5−10% range.[13]

What, then, is the marginal dead-weight loss from income taxation? As the discussion in Chapter 13 noted, no one knows for sure. The conventional wisdom is that the labor and capital market losses are likely to be the main leaks in Okun's bucket, but Feldstein's proposal to measure dead-weight loss by means of the elasticity of taxable income with respect to the after-tax rate is a powerful challenge to that wisdom (Feldstein, 1999). Also, the compliance leaks are large enough not to be ignored; they should figure prominently in any debate on tax reform, whether of the income tax or presumably of any other tax.[14]

The Shape of the Tax Schedule

The most unusual result with general tax schedules is that the marginal rates are not uniformly increasing throughout the range of income, in contrast to many actual tax schedules. Efraim Sadka was the first to demonstrate the by-now familiar result that the marginal tax rate at the top of the income scale should be zero (Sadka, 1976). This follows because the positive marginal tax rate at the top may reduce the labor supply of the highest income individuals. If the rate is dropped to zero and their labor supply increases, the government collects no revenue on this labor. But there was not any revenue on this marginal labor supply at the positive rate. So the only effect of setting the rate at zero is to increase the utility of the highest income individuals, which raises social welfare.

J. K. Seade demonstrated a similar result for the lowest incomes (Seade, 1977), namely, that as long as everyone who faces a positive wage chooses to work, the optimal marginal rate for the lowest income level is also zero. This follows because the only reason to levy positive rates at any income level, given that inefficiency will arise, is to redistribute the revenue to people below that income level. But no one is below the lowest income level. Hence, there is only an efficiency loss from taxing that income. Combining the Sadka and Seade results, the optimal general tax schedule must have a segment of rising marginal rates near the bottom and a segment of falling marginal rates near the top, contrary to the standard practice.

Another result of interest is the emerging consensus that not too much can be gained distributionally by a schedule of graduated tax rates relative to the linear income tax. This has practical significance because a flat-rate tax has much lower administrative and compliance costs. For example, it avoids the incentive to engage in tax arbitrage across tax brackets and the need to income average when incomes fluctuate over time.[15]

A U-Shaped Tax Schedule?

Diamond more recently made an important contribution to the optimal income tax literature. He showed that if the government chooses a nonlinear tax schedule, then the optimal pattern of marginal tax rates could well be

13. Slemrod (1992). An excellent review of the literature on tax compliance is Andreoni et al. (1998).

14. The Tax Reform Act of 1986 (TRA86) was the largest reform of the federal personal income tax ever undertaken, and it included some reform of the corporation income tax as well. It served as a large natural tax experiment for which the efficiency and equity effects have been intensely studied. Three excellent surveys of TRA86 are "Symposium on Tax Reform," *Journal of Economic Perspectives*, Summer 1987; "Symposium on the Tax Reform Act of 1986," *Journal of Economic Perspectives*, Winter 1992 (Auerbach and Slemrod, 1997).

15. An excellent overview of the income tax literature is provided in Slemrod (1983). In addition to the results reported here, Slemrod considers various extensions such as uncertain incomes, for which marginal tax rates have an added gain of providing social insurance against income losses; endogenous labor supply, in which taxation can lead to before-tax wage changes that imply a marginal subsidy on the highest income; and optimal income taxation in a dynamic framework that brings into play factors such as the treatment of future generations in the social welfare function and the Golden Rule of Accumulation, in addition to the intertemporal elasticity of substitution in consumption.

U-shaped, with marginal tax rates falling in a region below the modal level of skills and then rising in the region above the modal level. He also demonstrated that the optimal tax rates should probably continue to rise right up to the highest skill level, when they are then dropped to zero. That is, the result that the top marginal tax rate should be zero is of little practical importance (Diamond, 1998).

Diamond's demonstration of the likelihood of U-shaped marginal tax rates is based on a decomposition of the first-order conditions of the Mirrlees model that highlights the three main factors that determine the optimal pattern of marginal tax rates:

1. The compensated elasticity of the labor supply with respect to the wage (skill level), along with the probability density function at a given skill level and the skill level itself: These elements combine to determine the dead-weight loss from raising the marginal tax rate on an individual with a given skill level.
2. The difference between the social marginal utility of an additional dollar of government revenue and each individual's social marginal utility of income: This difference determines the social marginal benefit of increasing the tax rate on each individual.
3. The number of people with skills higher than the skill level on which the marginal tax rate is being raised. For the people with higher skills, the increase in the marginal tax rate is an inframarginal event; it affects their supply of labor only through income effects. Since income effects on labor supply are likely to be negative, it can be expected to increase the taxes they pay. Therefore, it has the potential of raising revenue from the higher skilled individuals without increasing efficiency loss.

Diamond's main contribution is in showing the effect of the distribution of skills on the optimal pattern of tax rates. To highlight the effect of the skills distribution, Diamond assumes that utility has the quasi-linear form $U = x + v(1 - y)$, where x is consumption, y is the supply of labor, and v is a concave function. With utility linear in x, the supply of labor is independent of income. This has two important implications for the optimal pattern of marginal tax rates. First, changes in the marginal tax rates on lower skilled individuals have no effect on the labor supply of the higher skilled individuals. The inframarginal effect on the higher skilled individuals is simply to raise revenue from them; there is no increase in efficiency loss from raising this revenue. Further, a per-unit lump-sum subsidy given to everyone also would have no effect on labor supply. This has the effect of setting the social marginal utility of an additional dollar of government revenue equal to the average value of the individuals' social marginal utilities of income. To simplify the analysis further for the sake of intuition about the distribution of skills, Diamond assumes a form of v that implies a constant labor supply elasticity.

Under these assumptions, consider the skill level of the person who has the average social marginal utility of income, which Diamond calls the critical skill level. For all people above that skill level, the difference between the marginal value of resources to the government and an individual's marginal utility of income is positive and ever increasing, assuming diminishing social marginal utility of income. Consequently, as skill levels increase, the average difference between the marginal value of resources to the government and individuals' social marginal utility of income over all the people at or above a given skill level continually increases as skills increase. This factor alone calls for steadily increasing marginal tax rates as skills increase. But the second factor that determines the pattern of tax rates is the ratio $(1 - F(n))/nf(n)$, that is, $1/n$ times the ratio of people with skill levels higher than n to the people with skill level n. Between the critical skill level and the modal skill level, this ratio is rapidly falling, sharply enough that it overrides the first factor and leads to falling marginal tax rates. The advantages of taxing the lower skilled people at higher marginal rates in that range are twofold. First, the government can raise proportionately more revenue from the inframarginal higher skilled individuals with no efficiency loss. Second, the direct efficiency loss of a given marginal tax rate is lower at lower n and lower $f(n)$.

Above the mode, Diamond assumes in one of his examples that the distribution of skills is the pareto distribution, for which $(1 - F(n))/nf(n)$ is constant. Only the first factor is relevant, and it implies that the marginal tax rates should be increasing. Therefore, the pattern of marginal tax rates is U shaped above the critical skill level, with the lowest marginal tax rate at or just above the modal skill level.

Diamond further shows that the marginal tax rates continue to increase at very high skill levels, so that the optimal switchover to a zero marginal tax rate at the highest skill level is likely to occur sharply at or near that skill level.[16] Moreover, the marginal tax rates near the top of the distribution can be very high, above 50% for some plausible values of the compensated labor supply elasticity, the distribution of marginal social welfare weights, and the pareto distribution over the range of high incomes in the United States.

Diamond's demonstration that the optimal pattern of marginal tax rates could be U shaped is particularly relevant in the United States because of the earned income tax credit (EITC). The EITC offers a wage subsidy to the working poor. In 2013, a single parent with two children received a

16. Diamond also shows that at two points in the distribution at which $f(n)$ is equal, one above the modal skill and one below the modal skill, the marginal tax rate should be higher for the lower skilled individuals. This turns out to happen because the factor $(1 - F(n))/nf(n)$ is sufficiently larger at the lower skill level to overcome the higher value of the first factor at the higher skill level.

40% subsidy on wage and salary income to $13,400, for a maximum credit of $5372. To avoid a severe drop-off of income plus subsidy immediately above $13,400, the credit remained constant at $5372 from earned incomes of $13,401 to $17,750. Then the credit decreased by 21 cents for each additional dollar of earned income until it became zero, at an income of $43,000, an additional effective marginal tax rate of 21%. Since a single-parent taxpayer was in the 15% tax bracket from $17,500 to $36,300, the effective marginal tax rate in that range was 36% (15% plus 21%). Then the tax rate jumped to 25% from $36,300 on, generating a marginal tax rate of 46% (25% plus 21%) from $36,300 to $43,000, the highest marginal tax rate faced by any taxpayer (the marginal tax rate was 39.6% for taxpayers with incomes above $450,000). Therefore, the marginal tax rate was U-shaped beyond $36,300, first 46%, then 25% and eventually rising at higher incomes as taxpayers moved into higher tax brackets. And it was essentially U-shaped from $17,500 on.

Diamond's analysis suggests that this type of pattern may be close to optimal. Still, a *U*-shaped tax schedule violates the spirit of Feldstein's no reversals principle, if only on the margin.

Concluding Observations

A caveat to all the results reported here is that an income tax by itself is not the optimal way to raise revenues unless preferences have certain separability properties that they are unlikely to have. Sufficient separability conditions for the optimality of a general income tax were given in the preceding section and in Chapter 14. Atkinson and Stiglitz also showed that a linear income tax is optimal if preferences are additively separable and the marginal disutility of labor is constant, very strong conditions indeed (See Atkinson and Stiglitz (1976)).

A final point worth noting is a methodology proposed by Erik Plug, Bernard van Praag, and Joop Hartog (PPH) for taxing ability, if a society were serious about trying to do this (Plug et al., 1999). In 1993, they surveyed people in the Noord−Brabant province of Holland who had taken intelligence quotient (IQ) tests as sixth graders in 1952. The idea was to regress current income on the 1952 IQ scores to obtain an estimate of earnings capacity (they also included years of education in the estimating equation in one version). They also used the survey to determine people's attitudes about income by asking them what levels of after-tax incomes they would place into six categories, from very bad to very good. The responses can be thought of as measuring the utility they receive from the different levels of income. The mean of the six income levels (in logs) is then regressed on family size, IQ, and income (to account for attitude drift related to income). The purpose of this regression is to standardize the utility received from mean attitudinal income levels across the respondents so they are comparable.

The next step in the methodology is to propose a utility function whose arguments are the individual's earnings capacity after tax and standardized mean attitudinal utility income. PPH chose the Leyden welfare function of income (ability) for the utility function.[17]

The final step is to use the utility function to design tax schedules on earnings capacity according to standard ability-to-pay sacrifice principles. PPH chose four sacrifice principles: (1) absolute equality of utility, (2) equality of marginal utilities, (3) equal proportional sacrifice, and (4) equal absolute sacrifice.

The most surprising result of this exercise is that the differences between the respondents' actual and optimal tax payments were fairly small, with the biggest differences arising from the equal marginal utility principle. The main advantage from taxing abilities in this manner is to eliminate discrepancies among people with equal earning capacities.

The methodology of PPH is a reasonable way to proceed to design a tax on ability, and they suggest various ways of improving the estimates of earnings capacities. Even so, the idea seems politically infeasible. In our view, the main value of their exercise is to underscore the likelihood that existing broad-based taxes will always be distorting. One especially uncomfortable implication of their approach is that a person's lifetime tax liability is at least partly determined by an IQ test taken as a youngster, whatever other adjustments PPH propose for the equation that estimates earnings capacity.

TAX EVASION

Private information raises the possibility of tax evasion, a problem that plagues all the developed nations. The Internal Revenue Service (IRS) estimated in 1998 that $282.5 billion of the tax liability under all federal tax liabilities went uncollected, 15.5% of the tax liabilities; $166.4 billion of the total was from the personal income tax, 17.6% of personal income tax liabilities.[18]

17. The Leyden welfare function is a log-normal distribution function of the log difference between the individual's after-tax earnings and standardized mean attitudinal utility income, divided by the standard deviation of the attitudinal utility income over its six categories of incomes. The standard deviation is set equal to the sample average standard deviation and therefore assumed to be constant across individuals.

18. *Budget of the United States Government, Fiscal Year 2009* (Washington, DC.: Government Printing Office, 2008), Part Five: Historical Tables, Table 2.1; and Statement of Leonard E. Burman, Senior Fellow, the Urban Institute, Codirector, the Tax Policy Center, Research Professor, Georgetown Public Policy Institute, Before the Committee on the Budget, United States House of Representatives, On Waste, Fraud, and Abuse in Federal Mandatory Programs, July 9, 2003, Figure 1. Burman noted in his statement that these were only rough estimates, because the IRS stopped measuring noncompliance carefully in 1988, with the single exception of the noncompliance of the working poor.

Tax experts and economists distinguish between tax avoidance and tax evasion. Tax avoidance refers to taxpayers taking advantage of the provisions of the tax laws to reduce their tax liability, such as arranging to take income in the form of lightly taxed capital gains or untaxed fringe benefits rather than as fully taxed wages and salaries. Avoiding taxes is legal and its consequences are certain. Tax evasion refers to hiding sources of taxable income from the tax authorities to reduce one's tax liability, such as not reporting gambling winnings or failing to file a tax return when required to do so. Evading taxes is illegal and its consequences are uncertain; they depend on the probability of the taxpayer being caught.

Sometimes the line between avoidance and evasion is fuzzy given the complexities of the tax law. For example, a taxpayer may be unclear whether a certain expenditure is deductible under the personal income tax but takes the deduction anyway. Is this avoidance or evasion? In any event, the traditional distinction between the two—legal and certain (avoidance) versus illegal and uncertain (evasion)—will suffice for our purposes. We begin with the problem of tax evasion.

The standard economic model of tax evasion is essentially identical to the economic model of criminal activity. The dishonest tax evader weighs the gains from evasion in the form of higher after-tax income against the costs, which depend on the probability of being caught and the penalties for evasion. The following very simple model captures the essentials of the economics of tax evasion.

Suppose a taxpayer has a fixed income Y that is subject to a personal income tax at a constant rate t. If the taxpayer is honest, his after-tax income is $Y_{AT} = (1 - t)Y$. If the taxpayer is dishonest and chooses to hide some of his income to evade taxes, the probability of being caught is p and the penalty is a fine equal to s times the amount of the undeclared tax liability. The fine is in addition to the tax at rate t on the undeclared income. Suppose the dishonest taxpayer declares income of Y^D, with $Y^D < Y$. The possible outcomes are

If not caught: $Y'_{AT} = Y - tY^D$

If caught: $Y''_{AT} = Y - tY - st(Y - Y^D) = Y(1 - t - st) + st\ Y^D$.

The usual assumption in tax evasions models is that the dishonest taxpayers are Von Neuman−Morgenstern expected utility maximizers, with

$$E(U) = (1 - p)U(Y'_{AT}) + pU(Y''_{AT}) \qquad (15.47)$$

The dishonest taxpayer's problem is to determine the level of income to declare, Y^D, that maximizes $E(U)$, given the opportunity locus between being caught and not being caught.

The opportunity locus as a function of Y^D is depicted in Fig. 15.5, with Y''_{AT} (being caught) on the vertical axis and Y'_{AT} (not being caught) on the horizontal axis.[19] The 45° line is a convenient frame of reference. The opportunity locus is the line segment AB. Point A on the 45° line applies if the taxpayer is honest and declares all his income. Y_{AT} equals $Y(1 - t)$. Point B applies if the taxpayer declares none of his income. If he is not caught, $Y'_{AT} = Y$. If he is caught, $Y''_{AT} = Y(1 - t - st)$, because he pays the penalty rate s on the entire tax liability tY. The slope of AB is $-s$, which is immediately evident from the expressions for Y'_{AT} and Y''_{AT} above. An additional dollar of declared income Y^D reduces Y'_{AT} if not caught by t, and raises Y''_{AT} by st if caught. Therefore, the slope $dY''_{AT}/dY'_{AT} = (-)st/t = -s$ along AB.

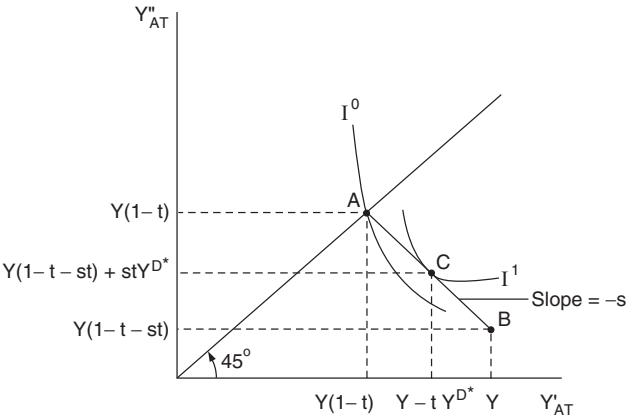

FIGURE 15.5

Regarding preferences, note that the slope of an expected utility indifference curve equals $(1 - p)/p$ at its intersection with the 45° line, with $Y''_{AT} = Y'_{AT}$.[20]

The dishonest taxpayer maximizes expected utility at point C, declares Y^{D*}, and winds up with either $Y'_{AT} = Y - tY^{D*}$ or $Y''_{AT} = Y(1 - t - st) + stY^{D*}$.

Readers familiar with the finance literature will notice that the analysis of tax evasion, and of criminal behavior generally, is closely related to the analysis of risk taking in finance. But there is one important difference. Risk under tax evasion is not exogenous. The government influences both s and p, the former directly and the latter indirectly by the efforts it takes to monitor taxpayers and enforce the tax laws. Consequently, two comparative static exercises of particular interest in the tax-evasion model are the effects on Y^D of changes in s and p. Both can be seen from Fig. 15.6.

19. From this point on, the analysis follows the presentation in Cowell (1985). Cowell's article is an excellent, wide-ranging review of the literature on tax evasion up to 1985 and offers insightful comments on a large number of issues surrounding tax evasion.

20. Along an indifference curve $dE(U) = 0 = (1 - p)U'(Y'_{AT})dY'_{AT} + pU'(Y''_{AT})dY''_{AT}$. $Y''_{AT} = Y'_{AT}$ along the 45° line, so that $(-)\ dY''_{AT}/dY'_{AT} = (1 - p)/p$.

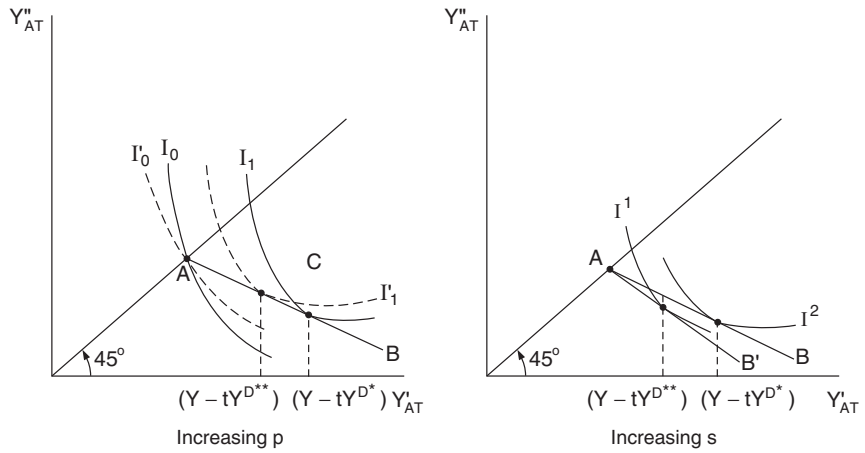

FIGURE 15.6

Increasing the Penalty

An increase in the penalty s rotates the opportunity locus around point A, making it steeper. Y^D increases, since Y''_{AT} and Y'_{AT} are both normal goods. Indeed, the incentive to hide income disappears entirely if s is raised to $(1-p)/p$, since the equilibrium would then be at A, with all income declared. This is a common result in the analysis of criminal activity: a high enough penalty deters all criminal activity providing utility is unbounded from below.

Increasing Monitoring

An increase in p flattens the indifference curves at their point of intersection with the 45° line and therefore everywhere along the curves, assuming continuity. Once again Y^D increases. An increase in monitoring activity reduces the incentive to hide income, as expected. Raising p such that $(1-p)/p = s$ removes all incentive to hide income.

The tax authorities thus have two effective methods of deterring tax evasion, but they are not equivalent. Increasing monitoring and enforcement efforts to increase p are likely to be far more expensive than increasing the penalty s, especially if s simply involves a fine and not incarceration. Society might not be willing to increase s high enough to reduce tax evasion, however. We have seen that s must be at least as large as $(1-p)/p$ to eliminate evasion. Suppose p is small, on the order of 0.10. Then the penalty has to be nine times the tax liability, a hefty fine indeed. The legal principle that "the penalty must fit the crime" might explain why tax authorities (and other policing authorities) engage in costly monitoring and enforcement efforts in lieu of imposing extremely harsh penalties.

A final comment on the model of tax evasion is that the assumption of fixed income does not bias the results in an important manner. A number of models with endogenous income exist in the literature. A common strategy is to tie evasion to the labor supply decision, in which people decide how much to work in the regular economy and how much in the underground economy where their labor income is hidden from the tax authorities. The analysis typically includes a social welfare function and attempts to determine the social welfare maximizing levels of monitoring and enforcement activities or penalties.

Models that assume a utilitarian social welfare function can produce a counterintuitive result that reducing enforcement activity may be social welfare enhancing. The reason is that the dishonest taxpayers are given the same weight as the honest taxpayers under utilitarianism, and reduced enforcement makes the dishonest taxpayers better off. If dishonesty is widespread, social welfare increases with reductions in enforcement. This result underscores the problem that private information poses for a normative public sector theory, which was mentioned in the introduction to the chapter: It can make the choice of appropriate norms, the social objective function, somewhat problematic. Why, exactly, does society want to reduce tax evasion? Mainstream normative public sector models do not give us a clear-cut answer.

Revenue-Raising Strategies

Increasing monitoring and enforcement activities and increasing penalties are not just means to reduce illegal activity. They also have the effect of raising tax revenues, and in this regard they become an alternative to raising tax rates. Joel Slemrod developed a simple and intuitive model to analyze how the interaction of these different revenue-raising strategies affects social welfare (Slemrod, 1994).

Slemrod posits the standard two-class society consisting of one (representative) high-ability person and one

low-ability person who are otherwise identical. The high-ability person receives a wage W_H and the low-ability person a wage of W_L, which are also their incomes because the supply of labor is fixed at one. The government's principal activity is a tax-transfer redistribution from the high-ability person to the low-ability person to maximize an Atkinson social welfare function. Secondarily, it also expends an amount E on enforcement activity to prevent the high-ability taxpayers from evading taxes. The high-ability person is assumed to have private information about his or her income that allows hiding any amount of income from the tax authorities that he or she wishes.

The driving element of Slemrod's model is a cost function that the high-ability person faces if he or she chooses to evade taxes. The cost function has the form

$$C = 1/2\left(EA^2/a\right) \tag{15.48}$$

where

$E =$ enforcement expenditures by the government.

$A =$ the amount of income hidden from taxation.

$a =$ a technological parameter that represents the ease of avoiding or evading taxes.

Slemrod refers to A as tax avoidance, but it could be either avoidance or evasion.[21]

Consider first the high-ability person. The person's only economic decision in this simple model is to determine the amount of income to hide from the government to maximize the after-tax income, given a, E, and t:

$$\max_{(A)} Y_{AT}^{H} = W_H - t(W_H - A) - 1/2(EA^2/a)$$

The first-order condition for A is

$$t - EA/a = 0, \text{or} \tag{15.49}$$

$$A^* = \text{at}/E \tag{15.50}$$

Therefore,

$$\begin{aligned} Y_{AT}^{H*} &= \left(1-t\right)W_H + at^2/E - 1/2\left(at^2/E\right) \\ &= \left(1-t\right)W_H + 1/2\left(at^2/E\right) \end{aligned} \tag{15.51}$$

Now consider the government's decision. The government budget constraint is

$$R = t(W_H - A) - E \tag{15.52}$$

with all the R transferred to the low-ability person. The government's objective is to maximize an Atkinson social welfare function with respect to t and E, given the high-ability person's optimal response to any given t and E. Formally,

$$\begin{aligned} W &= 1/\alpha\left\{\left[Y_{AT}^{H*}\right]^{\alpha} + \left[Y_{AT}^{L}\right]^{\alpha}\right\} \\ &= 1/\alpha\left\{\left[\left(1-t\right)W_H + 1/2\left(at^2/E\right)\right]^{\alpha}\right. \\ &\qquad \left. + \left[W_L + t(W_H - at/E) - E\right]^{\alpha}\right\} \end{aligned}$$

Slemrod simulates the model for $W_H = 3$, $W_L = 1$; $\alpha = -1, -2, -3$; and $a = 0.5$, 1.0, and 1.5.

A result of particular interest is the response of t and E to different values of a, given α. Technological change relating to the ability to monitor and hide income has been an increasingly important factor in tax policy since the 1980s. Computer technology has greatly enhanced the IRS's ability to monitor income, but the monitoring gains have undoubtedly been more than offset by a number of developments in financial markets that have made it much easier to hide income from capital. These include the Monetary Decontrol Act of 1980, which broke down most of the regulatory barriers in financial markets, coupled with that same computer technology that internationalized financial markets and facilitated the development of many new kinds of sophisticated assets and liabilities. For one thing, it is now much easier to move income "offshore" to escape taxation.

The net effect of these changes can be viewed as an increase in the technological parameter a in Slemrod's simple model, which lowers the cost of evasion. For $\alpha = -1$, the most realistic of the three aversion-to-inequality values for the United States, Slemrod found that t should decrease and E should increase. For $\alpha = -3$, however, he found that both t and E should increase. t decreases only if E is held constant.

Slemrod also conducted simulations in which he imbedded the avoidance cost function and his government budget constraint in the Mirrlees optimal income tax problem. The findings on t and E were much the same as in his simple model. In particular, if avoidance (evasion) is concentrated among the high-income people, as is likely, then the optimal t is lower for a given E because the tax rate is a less effective redistributive instrument. But the elasticity of avoidance with respect to the tax rate has an important role to play. A higher elasticity implies a lower t, as expected. At the same time, however, a higher E lowers the elasticity, which implies a higher t. The point is that t cannot be set independently of E.

These results led Slemrod to an interesting observation about the dramatic tax policy of the 1980s. The Tax Code was significantly amended twice in the 1980s, once in 1981 and again in 1986, with the result that the marginal tax rate on the highest income rates fell from 70% in 1980 to 28% in 1986. The decline in the highest marginal tax rate was widely viewed by economists as a triumph for the

21. Avoidance is costly because of the record keeping required on deductible items and either the time spent learning the tax laws or the fees paid to a tax preparer.

optimal income tax model. The consensus result from the model at the time was that the highest marginal tax rate should be on the order of 30%. These models incorporated only the dead-weight losses associated with labor supply and savings, however. Research that incorporated the administrative and compliance costs in the face of private information was yet to come. Slemrod observes, based on his model, that a sharp decrease in the tax rate was called for, given the increasing ease of avoidance/evasion (the increase in a) and also that monitoring and enforcement activity (E) was not increased. But if E had been adjusted optimally, then perhaps t should not have been cut so drastically. The increase in the highest marginal tax rate to 39.6% during the Clinton administration might have been called for, providing monitoring and enforcement efforts were also increased. The return to the highest 39.6% marginal tax rate in 2013 under the Obama administration could not be justified on the same grounds, however. The Republicans controlled the House of Representatives at that time and cut the IRS budget, to the point that the IRS had to sharply curtail its monitoring and enforcement efforts. Whatever the reasons for raising the top tax rates in these two instances, the new lesson in tax theory is that the option of increasing revenues through increased monitoring and enforcement should not be ignored.

Tax Amnesties

Tax amnesties are an attempt by the tax authorities to deal with the problem of tax evasion after the fact. The Department of Revenue declares an amnesty period of a few months in which taxpayers can declare previously hidden income and pay taxes on it without an additional penalty. Tax amnesties have been popular with state governments; the states declared 34 amnesties from 1981 to 1992.

The effectiveness of tax amnesties has been the subject of sharp debate. Those in favor of amnesties believe that they bring taxpayers out of the cold and turn them into law-abiding taxpayers from then on. According to this view, amnesties will reduce future tax evasion. Those opposed to amnesties claim that they are likely to backfire among the honest taxpayers, who will resent the break given to the dishonest taxpayers. Even worse, the honest taxpayers will realize how widespread tax evasion is and, given their resentment, be more prone to cheating themselves. The net effect will be an increase in tax evasion.

James Alm and William Beck used times-series econometric techniques to test the effects of a tax amnesty in Colorado in 1985 that ran from September 15 to November 15, 1985. They analyzed monthly state income revenue collections from January 1980 through December 1989 and found that the amnesty had no effect on monthly revenue collections whatsoever, neither after the amnesty nor even during the amnesty period. Perhaps the incentives noted above for amnesties to decrease and increase dishonesty essentially cancel one another, although this is pure conjecture (Alm and Beck, 1993).

CONCLUDING REMARKS

The formal second-best analysis of taxation that combines the dual concerns for efficiency and equity dates from the 1960s. It has gone through two distinct stages. The first stage is represented by Chapters 13 and 14. It explored optimal taxation under the assumption that the tax instruments chosen by the government were fixed and that monitoring and enforcement were not an issue. The government may or may not levy lump-sum taxes; it may tax virtually everything or only a restricted subset of goods and factors. Whatever the government chooses to tax, however, it knows exactly how much revenue it will collect. The second stage, beginning in the early 1970s, introduced private information into the formal analysis, which has been the subject of this chapter. The concerns about the effects of private information have more or less won the day. In a 1990 article on optimal taxation, Joel Slemrod noted that the leading questions surrounding tax policy now all turn importantly on the problem that taxpayers have private information. He mentions three of the more important ones (Slemrod, 1990).

The first is whether the government should tax consumption or income, in terms of which tax is easier to administer. Most of the complications of an income tax are associated with the taxation of income from capital (saving), which is avoided under a consumption tax. Is it really possible, for example, to tax all income from capital at the same rate, when some of the returns take the form of unrealized capital gains or in-kind services from real assets such as houses or rare paintings? Whether a consumption tax is really simpler is unclear, however. Taxpayers would have to register all their savings to verify the deduction taken from income in determining their tax liability, and this may not be straightforward either. The only truly simple tax on individuals to administer is a tax just on wage and salary income. Taxpayers could then report their labor income and their resulting tax liability on a single postcard. The wage tax has been the long-standing proposal of Robert Hall and Alvin Rabushka, dating from 1983 with the publication of their monograph *Low Tax, Simple Tax, Flat Tax*.[22] Whether the United States is willing to forego taxing income from capital remains an open question, however, as it is with the consumption tax.

22. Hall and Rabushka (1983). More recently, Hall and Rabushka (1995). Recall from the discussion in Chapter 11 that wage and consumption taxes are approximately equivalent in a static context, but not in a dynamic overlapping generations (OLG) context.

A second issue, assuming Congress retains the income tax, is whether the graduated rates should be replaced by a flat rate. On the one hand, the flat rate would be less redistributive beyond the low end of the income distribution. Low-income families and individuals would presumably still be protected from taxation by an initial exemption. On the other hand, the flat rate would remove the incentives for tax arbitrage that exist under graduated rates. Roughly speaking, graduated rates give high-income taxpayers an incentive to be long in (own) lightly taxed assets and be short in (borrow) heavily taxed assets even if the assets have the same risk and return characteristics.

A final issue is the one addressed earlier, whether the government should increase its revenues by increasing tax rates or beefing up its monitoring and enforcement efforts.

In summary, the new issues are related to the search for the optimal tax system in the presence of private information rather than the optimal set of tax rates within an assumed tax system. They recognize that the efficiency costs of taxation are much larger than the dead-weight losses associated with labor supply and saving that were the focus of the first-stage analysis. Administration and compliance costs may not be as large as the market distortions, but they are large enough to force tax experts to think about what kinds of taxes ought to be used. The tax system should not be taken as a given in normative tax analysis.

REFERENCES

Alm, J., Beck, W., March 1993. Tax amnesties and compliance in the long run: a time series analysis. National Tax Journal Vol. 48 (1), 53–60.

Andreoni, J., Erard, B., Feinstein, J., June 1998. Tax compliance. Journal of Economic Literature Vol. 36 (2), 818–860.

Atkinson, A., 1973. How progressive should income tax be? In: Parkin (Ed.), Essays in Modern Economics. Longman Group, Ltd., London.

Atkinson, A., Stiglitz, J., July/August 1976. The design of tax structure: direct versus indirect taxation. Journal of Public Economics Vol. 6 (1-2), 55–75.

Auerbach, A., Slemrod, J., June 1997. The economic effects of the tax Reform act of 1986. Journal of Economic Literature Vol. 35 (2), 589–632.

Boadway, R., Marchaud, M., Pestieu, P., September 1994. Towards a theory of the direct-indirect tax mix. Journal of Public Economics Vol. 55 (1), 71–88.

Cowell, F., September 1985. The economic analysis of tax evasion. Bulletin of Economic Research Vol. 37 (3), 163–193.

Diamond, P., March 1998. Optimal income taxation: an example with a U-shaped pattern of optimal marginal tax rates. American Economic Review Vol. 88 (1), 83–95.

Feldstein, M., November 1999. Tax avoidance and the deadweight loss on the income tax. Review of Economics and Statistics Vol. 81 (4), 674–680.

Hall, R., Rabushka, A., 1983. Low Tax, Simple Tax, Flat Tax. McGraw-Hill, New York.

Hall, R., Rabushka, A., 1995. The Flat Tax. Hoover Institution Press, Stanford, CA.

Hoff, K., Lyon, A., November 1995. Non-leaky buckets: optimal redistributive taxation and Agency costs. Journal of Public Economics Vol. 58 (3), 365–390.

McLaren, J., May 1998. Black markets and optimal evadable taxation. Economic Journal Vol. 108, 665–679.

Mirrlees, J., April 1971. An exploration in the theory of optimum income taxation. Review of Economic Studies Vol. 38 (2), 175–208.

Plug, E., van Praag, B., Hartog, J., May 1999. If we knew ability, how would we tax individuals? Journal of Public Economics Vol. 72 (2), 183–211.

Sadka, E., June 1976. On income distribution, incentive, effects, and optimal income taxation. Review of Economic Studies Vol. 43 (2), 261–268.

Sah, R., February 1983. How much redistribution is possible through commodity taxes? Journal of Public Economics Vol. 20 (1), 89–101.

Seade, J., April 1977. On the shape of optimal tax schedules. Journal of Public Economics Vol. 7 (2), 203–235.

Slemrod, J., September 1983. Do we know how progressive the income tax system should Be? National Tax Journal Vol. 36 (3), 361–369.

Slemrod, J., Winter 1990. Optimal taxation and optimal tax systems. Journal of Economic Perspectives vol. 4 (1), 157–178.

Slemrod, J., Winter 1992. Did the tax Reform Act of 1986 simplify tax matters. Journal of Economic Perspectives Vol. 6 (1), 45–57.

Slemrod, J., September 1994. Fixing the leak in Okun's bucket: optimal tax progressivity when avoidance can be controlled. Journal of Public Economics Vol. 55 (1), 41–51.

Stern, N., July/August 1976. On the specification of models of optimum income taxation. Journal of Public Economics Vol. 6 (1-2), 123–162.

Stiglitz, J., Pareto efficient and optimal taxation and the new welfare economics, In: Auerbach, A., Feldstein, M., (Eds.), Handbook of Public Economics, vol. II, Elsevier Sciences Publishers B. V., North-Holland, Amsterdam (chapter 15).

Symposium on Tax Reform, *Journal of Economic Perspectives*, Summer 1987, Vol. 1 (1), 3–119.

Symposium on the Tax Reform Act of 1986, *Journal of Economic Perspectives*, Winter 1992, Vol. 6 (1), 3–68.

Chapter 16

The Theory and Measurement of Tax Incidence

Chapter Outline

When a tax is levied on an economic agent, the agent is said to bear the *impact* of the tax, equal to the amount of the tax payment. Economists distinguish the impact of a tax from the *incidence* or burden of a tax. The distinction is important because the pattern of tax payments may not be a very good measure of the true economic burdens arising from a tax. The problem is that a tax initiates, potentially, an entire chain of general equilibrium market effects that can change consumer and producer prices. These price changes, in turn, generate welfare losses and gains throughout the economy that affect, potentially, all economic agents, not just those who paid the tax. The incidence or burden of a tax incorporates both the initial impact of the tax and the gains and losses associated with the general equilibrium market reactions to the tax. As such, the incidence and not the impact of a tax is the central concept of interest in either a normative or a positive distributional theory of taxation.

TAX INCIDENCE: A PARTIAL EQUILIBRIUM ANALYSIS

All students of economics are introduced to the distinction between the impact and incidence of a tax at the principles level, at least in a partial equilibrium context. Recall the

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00016-9

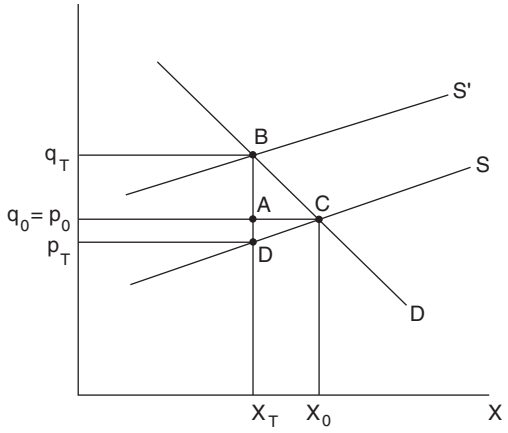

FIGURE 16.1

standard analysis of a unit sales tax paid by all producers in a competitive market, depicted in Fig. 16.1. The unit tax shifts the supply curve up vertically by the amount of the tax, because each producer's marginal cost at any given output rises by the amount of the tax. The shift in the supply curve can be thought of as the suppliers' attempt to pass the tax on to the consumers through higher prices. Whether or not they succeed depends upon the elasticities of both supply and demand. As drawn in Fig. 16.1, the price to the consumer rises, but only from q_0 to q_T, less than the full amount of the tax. The producer price falls from p_0 to p_T ($=q_T - t$). The new equilibrium output is X_T, and the tax revenue is $X_T \cdot t = X_T(q_T - p_T)$.

The impact of the tax falls on the producers and is equal to the total tax payment, but the incidence or true burden of the tax is shared by the producers and consumers. Because the consumer price (the "gross-of-tax" price) rises from q_0 to q_T, the consumers suffer a loss of consumer surplus (ignoring income effects) equal to $q_T BC q_0$. Because the producer price (the "net-of-tax" price) falls from p_0 to p_T, the producers suffer a loss of producer surplus equal to $p_0 CD p_T$. Using the impact, or tax revenue, as a measure of the true burdens, then, would overstate the producers' true economic losses by ($q_T BA p_0 - CAD$) and understate the consumer's true economic losses by $q_T BC q_0$.[1]

Even though this example is only a partial equilibrium analysis of tax incidence, it illustrates a fundamental point: The market's reactions to a tax are a crucial determinant of its ultimate pattern of burdens.

1. This example is meant only as an illustration of the distinction between tax impact and tax incidence. As we have noted in other contexts, partial equilibrium Marshallian consumer and producer surpluses are generally not valid measures of consumer and producer losses.

FIRST-BEST THEORY, SECOND-BEST THEORY, AND TAX INCIDENCE

We have delayed discussing the general theory of tax incidence until this part of the chapter, since the most interesting questions in tax incidence are inherently second best in nature, precisely because they depend on the market's response to distorting taxation. In first-best theory, all questions of distributional equity are incorporated into the interpersonal equity conditions,

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial I^h} = \frac{\partial W}{\partial U^j}\frac{\partial U^j}{\partial I^j} \quad \text{all } h,j = 1,\ldots,H \qquad (16.1)$$

which the government satisfies through a set of lump-sum taxes and transfers among the consumers. As a general rule, these lump-sum redistributions affect equilibrium market prices, contrary to the common assumption that they do not, and these price changes in turn, affect people's utilities. The incidence, or burden, on the consumers could presumably be measured as the difference in their utility levels before and after government redistribution, but computing the change in utility for each person is not especially interesting. Consider the various possibilities.

On the one hand, the equilibrium market prices may *not* change if, for example, aggregate production technology is linear with constant marginal opportunity costs. If producer prices remain unchanged, so too will the vector of consumer prices with the lump-sum redistributions. In this case, the tax and transfer payments are perfect income proxies for the change in utility in this sense—if any one person received (paid) the tax (transfer) back from (to) the government, his original utility level would be restored. Therefore, the impact and incidence of the redistribution are identical, so that the incidence question is trivial. In practice, tax theorists have often been willing to define the incidence of a tax (transfer) solely in terms of the distribution of tax payments (transfer receipts) if they believe the tax (transfer) is approximately lump sum. For example, many incidence studies of the personal income tax commonly allocate the tax on the basis of the tax payments by income class.

On the other hand, equilibrium prices do change in response to a lump-sum redistribution under general technology, the more realistic case. The redistribution moves society along the production–possibilities frontier as well as the utility–possibilities frontier. Now if the government restores the original income level for any *one* person, that person will not be able to achieve his original utility level, in general. The tax payment (transfer receipt) is not necessarily an accurate proxy for the change in utility.

Even so, the pattern of incidence is still not an especially compelling question, for two reasons. In the first place, the relevant alternative to a given program of lump-sum taxes and transfers can always be viewed as the

complete unraveling of the redistribution, in which the government restores everyone's original income level *simultaneously*. If this were done, the original general equilibrium and utility levels would also be restored. Under this assumption, then, the payments (receipts) by any one person can be viewed as a perfect proxy for the pattern of burdens since the price changes are irrelevant. Impact and incidence are again identical.

Suppose, however, that one insisted on viewing the problem strictly at the individual level, asking what the consequences would be to some individual of restoring his original income level while leaving the remaining taxes and transfers intact. The utility effects of the price changes are still not very interesting. Presumably the government's redistribution has taken account of the price-induced effects on individual marginal utilities in reaching the final equilibrium, at which the marginal social utilities of income are equalized. That the actual taxes and transfers may not be good proxies for any individual's utility gain or loss is really beside the point, because whatever changes in utility have occurred are the optimal changes required for a first-best social welfare maximum. There is no compelling reason to measure the incidence of the redistributions from a normative perspective, because no other pattern of redistributions could possibly dominate the given redistributions in the sense of being more equitable.

The same cannot be said for second-best taxes. In a second-best environment, taxes are raised in a distorting manner to meet certain revenue requirements. Particular taxes may be chosen simply on the basis of convenience or by some efficiency criterion such as minimizing deadweight loss, in which case it may well be possible to design more (or less) equitable taxes. If so, then measuring the incidence of the tax is important. Only if the taxes are optimally designed to maximize social welfare in accordance with an equation such as Eqn (14.38) would the question of tax incidence be more or less irrelevant, as it is in a first-best environment. But equations such as Eqn (14.38) are unlikely to hold in practice. Also, there is no most preferred second-best social welfare optimum because the optimum depends on the underlying constraints that make the environment second best. Therefore, economists have a clear motivation for developing accurate measures of incidence in a second-best environment.

In fact, the theory and measurement of tax incidence have been a central focus of public sector economics since the very beginning of the discipline, with the result that there exists a voluminous literature on the subject. The incidence of every major tax has been studied in detail, both theoretically and empirically. Rather than addressing each tax separately, Chapter 16 discusses the fundamental methodological issues underlying the theory and measurement of tax incidence in a second-best environment, issues applicable to all taxes.

METHODOLOGICAL DIFFERENCES IN THE MEASUREMENT OF TAX INCIDENCE

The tax incidence literature is bound to be confusing to a beginner in public sector economics. Empirical studies of individual taxes are fraught with controversy, in part because the empirical analysis of tax incidence is inherently so difficult, but also because there exist serious methodological differences among experts in the field, on the appropriate theoretical approaches to the measurement of incidence.

By way of illustration, consider the incidence of the U.S. corporate income tax, which a large number of researchers have studied. Their results could not possibly be more divergent. They range all the way from Richard Musgrave's early finding that the tax is borne at least 100% by the consumers of corporate output, to Arnold Harberger's estimate that corporate stockholders almost certainly bear virtually the entire burden of the tax, to Joseph Stiglitz's conjecture that the tax may be nondistorting. To confuse matters further, Ann Friedlaender and Adolph Vandendorpe showed that the analytical framework that Harberger used to determine the incidence of the tax should have generated the result that no one bears a burden (Krzyzaniak and Musgrave, 1963; Harberger, 1962; Stiglitz, 1973; Friedlaender and Vandendorpe, 1976). This was an important qualification, because Harberger's model is frequently used to study the general equilibrium incidence of taxes.

The corporation income tax may be the most dramatic instance of empirical uncertainty with regard to tax incidence, but the incidence of most other important taxes has hardly been settled either. To give one other example, most public sector economists had long believed that local property taxes were at least mildly regressive. In the 1970s, a "new view" consensus emerged that the property tax is almost certainly progressive.[2]

Perhaps it is not surprising that empirical estimates of the incidence of any tax should vary considerably, given the nature of the problem. Empirical researchers must select what they think are the most important market reactions to the tax from a staggering set of possibilities, and methods of selection are bound to differ. Unfortunately, empirical tax incidence analysis appears not to be especially robust to assumptions made about sectors of the economy not explicitly under examination. Another confounding factor, mentioned earlier, is that researchers often employ different *theoretical* measures of incidence as a basis for their empirical work, and it is the theoretical differences that we wish to focus on here.

2. Feldstein (1974). Refer also to Aaron (1974); Musgrave (1974) and comments. Also, Aaron (1975).

THEORETICAL MEASURES OF TAX INCIDENCE

Three distinct theoretical measures of incidence commonly appear in the literature: incidence as impact, incidence as changes in certain relative prices, and incidence as changes in welfare.

Impact Equals Incidence

Some research merely reports the pattern of tax payments by income class and judges the equity of the tax on this basis alone, thereby equating the impact and incidence of the tax. As noted above, most incidence studies of the personal income tax employ this measure, on the (inappropriate) assumption that income taxes are essentially lump sum. For example, Joseph Pechman and Bernard Okner allocate personal income-tax burdens in this manner in their widely cited Brookings studies, *Who Bears the Tax Burden* and *Who Paid the Taxes, 1966–85*? (Pechman, 1985; Pechman and Okner, 1974), which appeared in 1974 and 1985.

Changes in Relative Prices

At the other end of the spectrum, a large group of incidence studies base their measures of incidence on changes in certain market prices in response to a tax. The change in the wage-rental ratio is the usual choice. Actual tax payments influence this incidence measure because their impact and size affect both the pattern of general equilibrium price changes and the degree to which they change, but the tax payments themselves are not a part of the final incidence measure. Arnold Harberger pioneered this approach in his 1962 classic, *The Incidence of the Corporation Income Tax* (Harberger, 1962) and numerous other tax theorists have followed his lead. Changes in relative prices are featured prominently in dynamic tax incidence studies within the context of growth models, whether the models employ the Ramsey representative consumer assumption (e.g., Martin Feldstein) or the overlapping generations (OLG) with life-cycle consumers (e.g., Larwence Kotlikoff and Alan Auerbach) (Feldstein, 1974; Auerbach and Kotlikoff, 1987). Harberger's paper has certainly been one of the most influential works on the incidence question.

Changes in Welfare

A third approach is to relate tax incidence to changes in welfare, measured either directly as changes in individual's utility levels or indirectly as compensated income changes using the expenditure (and profit) function. John Shoven and John Walley, who pioneered the use of static computable general equilibrium (CGE) models to study tax incidence, follow this approach. So do Don Fullerton and Diane Rogers in their dynamic CGE analysis of tax incidence. Kotlikoff and Auerbach also provide welfare measures of gains and losses in their dynamic OLG models of tax incidence.[3] All these studies also report changes in general equilibrium price ratios as an intermediate step, following the spirit of Harberger's analysis.

Can such different incidence measures be reconciled? In our opinion, they cannot be, at least not fully, since they view the problem of measuring the burden of taxation from different perspectives that are some respects irreconcilable. Moreover, there appears to be no clear-cut presumption in favor of any one of them, or some other candidate not currently in vogue. They each have their advantages and disadvantages, and choosing among them ultimately depends on the researcher's personal preferences. No consensus best model or method has emerged for measuring tax incidence.

General Principles of Tax Incidence

Despite differences in the way they measure tax incidence, nearly all tax theorists agree on two general principles. The first is that people ultimately bear the burden of taxation, so that any notion of burden must relate either directly or indirectly to individual utilities. The second is that tax incidence must be analyzed within a general equilibrium framework.

That individuals bear the burden of taxation is merely a specific application of the first principle in all of normative public sector economics, that the government's task is to promote the interests of its constituents. Thus, although the government may levy a corporation income or sales tax on General Motors (GM), the interesting question in tax incidence is not the harm done to GM as a legal entity, but rather which *people* bear the burden of the tax—GM stockholders, GM workers, consumers of GM products, other consumers, other stockholders, other workers, and so forth—and how much of a burden each of them suffers. This principle also implies that any measure of burden should incorporate each individual's own perception of the burden he or she suffers as a result of the tax. As always in public sector theory, individual preference is a fundamental datum for public sector decision-making.

Regarding the general equilibrium framework, the overwhelming majority of tax incidence models assumes a fully employed economy with competitive markets, although there has been some work done on tax incidence in the presence of noncompetitive markets and/or unemployed resources.[4] In keeping with the rest of the text, we

3. Shoven (1976). Also, Shoven and Whalley (1972); Fullerton and Rogers (1993).

4. Refer to Asimakopulos and Burbridge (1974); Kalecki (1937).

will adopt the full-employment, competitive-market assumptions unless otherwise stated.

That tax theorists insist on general equilibrium modeling is altogether appropriate, given that the final burden of a tax directly depends on the pattern of the general equilibrium market responses to the tax. At the same time, however, this poses some sticky conceptual problems for tax incidence measurement.

The Disposition of the Tax Revenues

In the first place, general equilibrium analysis challenges the notion that it is possible to consider unambiguously the incidence of a single tax. The heart of the matter is that the disposition of the tax revenue must be accounted for explicitly in a proper general equilibrium framework. Tax revenue collected by the government cannot simply disappear without continued repercussions throughout the economy. The government most likely will spend the revenues in some manner, but it could simply save them. In any case, the disposition of the revenues generates its own pattern of welfare gains and losses. In what sense, then, can "the incidence of *a* tax" have meaning as an isolated phenomenon within a general equilibrium, interdependent market economy? This question deserves some careful thought.

Save the Tax Revenues

Suppose one assumes that the government simply saves the tax revenues in an attempt to isolate the incidence of a particular tax. This would appear to be the spirit of the many income-tax studies that assume the taxes are lump sum and thereby equate the impact of the tax with its incidence, without reference to the disposition of the revenues. As noted above, however, if the vector of general equilibrium prices changes in response to the tax, then impact and incidence are not generally equivalent for any one individual. Moreover, if one also assumes continued full employment, as is common with incidence analysis, then at least one price must change.

Suppose, for purposes of illustration, one chooses the simple standard Investment Saving–Liquidity Preference Money Supply (IS–LM) model of macroeconomics with competitive factor markets, depicted in Fig. 16.2, to analyze the consequences of taxation when the tax revenues are saved. According to this model, the real rate of interest changes from the pre- to posttax equilibrium. The tax shifts the IS curve to IS′, resulting in excess supply in the goods market. As a consequence, the absolute price level declines, increasing real money balances and shifting the LM curve to LM′. The full-employment level of real income is restored, but r has dropped from r_0 to r_1. Thus, in this model, the final burden of the tax depends not only on the pattern of tax payments but also on the welfare

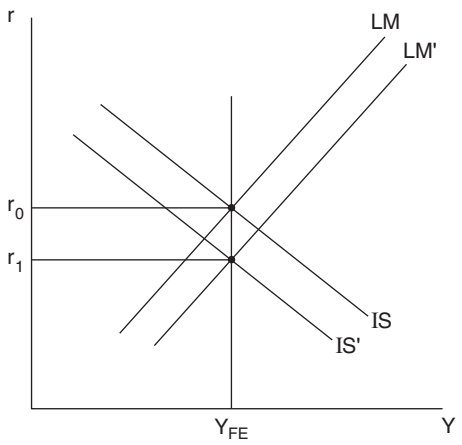

FIGURE 16.2

consequences of the decline in the real rate of interest. The impact and incidence of the tax are not identical, and the tax payments may be a poor proxy for the true pattern of welfare changes even when the tax revenues are saved.

Balanced Budget Incidence

Suppose, more realistically, that the tax revenue is used to finance government expenditures. The combined distributional effect of the tax and expenditure policy is commonly referred to as *balanced budget incidence*, which obviously depends upon the particular expenditures being financed. A number of possibilities exist. If the expenditures are lump sum, transfer payments and the taxes are also lump sum, then the tax-and-transfer program is first best, a case we have already considered. If the taxes are not lump sum, then the analysis is second best, even though the transfers happen to be lump sum. We will return to this point below.

If the expenditures take any other form, including distorting transfers or exhaustive expenditures, then the measurement of balanced budget incidence requires specific measures of the incidence of the expenditure programs as well as that of the taxes, a problem that we will consider in Chapter 17. Furthermore, these expenditure programs change the vector of equilibrium prices in general, which means that even a lump-sum tax payment may not be an accurate measure of the tax burden suffered by any individual consumer.

In conclusion, tax studies that use income-tax payments as the measure of incidence ought to be assuming that the taxes are lump sum *and* that they are being used to finance lump-sum transfer payments in a balanced budget manner (no shift in the IS curve). To be absolutely unambiguous, these studies should also assume a linear technology with unchanged producer prices. Then, as was also discussed above, if any *one* consumer received his tax payment back (returned his transfer), that payment would restore the consumer's original utility level. Otherwise, incidence and

impact are identical only in the aggregate sense described above, in which the alternative to the tax-and-transfer program is assumed to be a return to the original pretax and transfer equilibrium.

Pure Tax Incidence: Differential Incidence

Is there any way of focusing on the incidence of taxes per se in a general equilibrium framework without complicating the analysis with difficult questions of expenditure incidence? Theoretically the answer is yes, but one should keep in mind that taxes are usually changed in response to particular expenditure initiatives. Therefore the empirical relevance of pure tax incidence measures may be limited.

The Incidence of a Single Tax

One very popular method of analysis, initiated by Harberger, is to assume that the revenues collected are returned lump sum.[5] In a one-consumer or one-consumer-equivalent economy, there is no ambiguity over who receives the lump-sum returns. But the pattern of returns is crucial in a many-person economy, whether one considers the incidence effects from the aggregate viewpoint of social welfare or from each individual's perspective of the loss he suffers as a result of the tax with redistribution. The natural assumption for analytical purposes is that each person receives a lump-sum transfer exactly equal to his tax payment, so that the impact of the tax-and-transfer program on each individual is zero. While this assumption is surely a disservice to reality, it is a useful analytical device for considering the incidence of a single tax within the context of a general equilibrium framework. Assuming lump-sum returns of the revenues at least neutralizes the expenditure side of the budget as much as possible.

Differential Incidence

The other, more realistic, possibility is to consider the incidence of substituting one tax for another, while holding constant either total tax revenues or the entire government budget surplus (taxes—expenditures). This method of analysis is referred to as *differential incidence*, and since governments might actually do this, many researchers find it especially appealing. The method of returning tax collections lump sum to analyze the incidence of a single tax can be considered a specific case of differential incidence, in which the taxes being substituted for are head taxes levied on each individual consumer.

It is important to note that whether one chooses to hold tax revenues or the entire government budget surplus constant in a differential incidence analysis is a matter of some consequence. In order to focus strictly on differential tax incidence, the tax-revenue-constant assumption might appear to be the preferred alternative, but it may well violate the dictates of a full general equilibrium analysis. Suppose, for example, the government buys and sells goods and factors in the competitive-market system either at the producer or consumer prices. With general-technology production, both producer and consumer prices change in response to a tax substitution, thus changing both the level of government expenditures and the amount of revenue from the new tax necessary to balance the overall government budget. If tax collections are held constant in the process of substituting one tax for another, then the overall budget surplus may also change and this change has to be considered as part of the incidence analysis. Adding an assumption that government expenditures are also held constant simply poses different problems, for then government inputs and outputs must change, with corresponding changes in consumers' welfare. Only if the overall budget surplus (deficit) is held constant can the incidence analysis properly focus on the differential effects of the taxes. Government expenditures may change, but as long as the government does not vary the vector of government inputs and outputs, there can be no change in consumer(s) welfare arising from the expenditure side of the budget. Thus, the government-budget-surplus-constant assumption is preferred, even though total tax collections vary as one set of taxes is substituted for another.

The tax-revenue-constant and government-budget-surplus-constant assumptions are equivalent only if one assumes that linear production technologies exist in both private and public sectors *and* that government purchases (sales) are at producer prices. Since producer prices cannot change in response to taxation, neither does the level of government expenditures.

These considerations highlight the care that must be taken in order to specify a well-defined tax incidence analysis within a general equilibrium framework.[6] One must always specify what is being assumed about the disposition of the tax revenues and what effects the use of the revenues has on the general equilibrium of the economy.

Welfare Measures of Tax Incidence: One-Consumer Economy

Having determined that taxation with selective lump-sum return and, more generally, differential incidence are the

5. Harberger and others actually assume that the government spends the revenue exactly as the consumer(s) would have had they received it, but this is equivalent to redistributing the revenue lump sum and letting the consumer(s) spend it.

6. For an excellent general discussion of alternative equivalent taxes, see Shoven and Whalley (1977).

only appropriate methods for considering tax incidence independently from expenditure incidence in a general equilibrium framework, there remains the difficult theoretical problem of actually measuring the resulting incidence effects. Recall that three measures have been commonly employed: the impact of a tax, the change in (some) general equilibrium prices, and changes in utility or equivalent income compensation measures. The theoretical issues of measurement are sufficiently complex to warrant a preliminary discussion within the context of a one-consumer or one-consumer-equivalent economy, before turning to the more relevant many-person economy.

The first principle of tax incidence is that it should measure the burden on the consumer for any given pattern of taxation. This implies that the natural interpretation of burden in the one-consumer-equivalent economy is the deadweight loss measure developed in Chapter 13. For instance, if it is assumed that any tax revenue collected is simply returned to the consumer lump sum and that there are no government expenditures, the incidence of any given (set of) tax(es) would be appropriately measured as

$$L(\vec{\mathbf{t}}) = M\left(\vec{\mathbf{q}}; \overline{U}^0\right) - \sum_{i=1}^{N} t_i X_i^{\text{comp}} \quad \text{(linear technology)} \tag{16.2}$$

$$L(\vec{\mathbf{t}}) = M\left(\vec{\mathbf{q}}; \overline{U}^0\right) - \sum_{i=1}^{N} t_i X_i^{\text{comp}} - \pi(\vec{\mathbf{p}}) \quad \text{(general technology)} \tag{16.3}$$

identical to the measurement of deadweight loss from taxation. With linear technology, recall that $M(\vec{\mathbf{q}}; \overline{U}^0)$ measures the lump-sum income necessary to compensate the consumer for a given vector of taxes, $\vec{\mathbf{t}}$, with $\overline{U}^0$ equal to the zero-tax level of utility. With general technology, there may be pure profits or losses from production as producer prices, $\vec{\mathbf{p}}$, vary in response to the tax rates. In this case, the appropriate income measure of welfare loss is $M(\vec{\mathbf{q}}; \overline{U}^0) - \pi(\vec{\mathbf{p}})$, where $\pi(\vec{\mathbf{p}})$ is the general equilibrium profit function. With the tax revenues returned lump sum, Eqns (16.2) and (16.3) measure the consumer's loss.

Tax incidence and tax inefficiency are equivalent because, with the taxes returned lump sum, the *only* source of welfare loss is the change in the vector of consumer (and producer) prices resulting from the taxes. The tax payment, the impact of the tax, affects this measure only indirectly. The level of the tax rates, $\vec{\mathbf{t}}$, in part determines the amount by which the consumer (and producer) price vectors change, exactly as in loss measurement.

The loss measurement is also appropriate for the measurement of differential incidence. The substitution of one tax for another with revenue held constant can be thought of as follows: Impose one set of taxes and return the revenues lump sum. Then impose a different set of taxes with its revenues returned lump sum. In short, differential *incidence* is equivalent to differential *efficiency* in a one-consumer economy. It follows exactly the framework developed for the Corlett–Hague analysis presented in Chapter 13. Recall that the relevant equations for marginal changes in tax rates are in vector notation:

$$\mathrm{d}L = \frac{\partial L}{\partial t}\mathrm{d}t \tag{16.4}$$

and

$$\mathrm{d}T = 0 = \frac{\partial (tX)}{\partial t}\mathrm{d}t \tag{16.5}$$

Equation (16.5) determines the changes in tax rates necessary to maintain tax revenues constant, and Eqn (16.4) computes the resulting change in the consumer's welfare in terms of the lump-sum income required to hold utility constant.

In the case of general technologies, the market clearance equations relevant to the compensated equilibrium,

$$X_i\left(\vec{\mathbf{q}}; \overline{U}^0\right) = X_i\left(\vec{\mathbf{p}} + \vec{\mathbf{t}}; \overline{U}^0\right) = \pi_i(\vec{\mathbf{p}}) \quad i = 2, \ldots, N \tag{16.6}$$

are also necessary to relate market price changes to the tax changes in computing Eqn (16.4). (Recall that compensation is assumed to occur in terms of the numeraire good, the first good as written, which is also assumed to be untaxed. These assumptions have already been discussed in the development of the deadweight loss from taxation. For linear technologies, $\mathrm{d}q = \mathrm{d}t$ and $\mathrm{d}p = 0$, so that market clearance is unnecessary.) Note, finally, that the incidence or loss from lump-sum taxes is zero according to these measures, precisely because they entail zero deadweight loss. Clearly, collecting lump-sum taxes and returning the revenues lump sum cannot give rise to an economic burden in a one-consumer or one-consumer-equivalent economy.

The equivalence between incidence and deadweight loss from taxation carries over in the presence of government expenditures. Suppose the government's budget constraint is

$$\sum_{i=1}^{N} t_i X_i + \sum_{i=1}^{N} p_i Z_i = S \tag{16.7}$$

where

S = the fixed government surplus (possibly equal to zero or negative, a deficit) and
Z_i = the government purchase (supply) of input (good) i, with all government transactions at competitive producer prices.

As long as any government surplus, S, is returned to the consumer lump sum, an appropriate assumption for general equilibrium analysis, then the deadweight-loss measurements for any given pattern of government decision variables $(\vec{\mathbf{t}}; \vec{\mathbf{Z}})$ are

$$L(\vec{\mathbf{t}}; \vec{\mathbf{Z}}) = M\left(\vec{\mathbf{q}}; \overline{U}^0\right) - \sum_{i=1}^{N} t_i X_i - \sum_{i=1}^{N} p_i Z_i$$

$$\text{(linear technology)} \tag{16.8}$$

$$L(\vec{\mathbf{t}}; \vec{\mathbf{Z}}) = M\left(\vec{\mathbf{q}}; \overline{U}^0\right) - \sum_{i=1}^{N} t_i X_i - \sum_{i=1}^{N} p_i Z_i - \pi(\vec{\mathbf{p}})$$

$$\text{(general technology)} \tag{16.9}$$

For a constant value of Z, these are also the appropriate measures for the incidence of any given vector of taxes $\vec{\mathbf{t}}$ providing the government surplus is returned lump sum. Moreover, the equations (in vector notation)

$$\mathrm{d}L = \frac{\partial L}{\partial t}\mathrm{d}t \tag{16.10}$$

and

$$\mathrm{d}S = 0 = \frac{\partial(tX - pZ)\cdot \mathrm{d}t}{\partial t} \tag{16.11}$$

determine the differential incidence of any tax substitutions that leave the overall government budget surplus unchanged.[7] If the surplus is held constant, one can think of differential incidence as replacing one set of taxes having the surplus returned lump sum with another set of taxes also having the (same) surplus returned lump sum.

The Relative Price Measure of Differential Tax Incidence: One-Consumer Economy

Although the notion of income compensation provides a nice theoretical bridge between tax inefficiency as represented by deadweight loss and tax incidence, the loss measures may well have limited applicability to the practical requirement of deriving empirical measures of tax incidence. The problem, which was also addressed in Chapter 13, is that loss measures require knowledge of compensated equilibria and they are not observed in practice. This is not so serious with linear production technologies, since any pattern of tax rates generates the same set of producer and consumer prices at both the compensated and actual with-tax equilibria. Even so, the amount of tax revenues collected for any given set of rates differs between the two equilibria. Therefore, the incidence of a given set of taxes can be thought of as the incidence of establishing a given set of tax *rates* and then returning the resulting revenues lump sum. The loss measure is then unambiguously defined for the given tax rates, but not for a given amount of revenue. Differential incidence would be measured analogously. Presumably one would determine a set of tax rates as an alternative to a given set of tax rates that held *actual* tax collections (or the overall budget surplus) constant and then use those changes to compute Eqn (16.4) or (16.10). Since $\mathrm{d}q = \mathrm{d}t$ and $\mathrm{d}p = 0$ in both the actual and compensated equilibria, the loss can be evaluated unambiguously for the given pattern of $\mathrm{d}t$. Of course, the vector dt that keeps actual tax revenues constant does not, in general, hold compensated tax revenues constant, but it is still possible to mix compensated and actual equilibria in the manner suggested.

With general technologies, however, it is not clear how to use the loss measure. Consider the problem of measuring the incidence of a given set of tax rates when the tax revenue has been returned lump sum to the consumer. Presumably one wants to measure the loss implied by the given set of tax rates, $\vec{\mathbf{t}}$, and the market prices, $\vec{\mathbf{q}}_A$ and $\vec{\mathbf{p}}_A$, observed in the actual with-tax equilibrium. With general technologies, however, any given vector, $\vec{\mathbf{t}}$, generates one set of market prices $(\vec{\mathbf{q}}_A, \vec{\mathbf{p}}_A)$ in the actual equilibrium and a different set of prices $(\vec{\mathbf{q}}_c, \vec{\mathbf{p}}_c)$ in the compensated equilibrium. This follows because the market clearance equations for the actual and compensated equilibria,

$$X_i^{\text{actual}}(\vec{\mathbf{p}} + \vec{\mathbf{t}}) = \pi_i(\vec{\mathbf{p}}) \quad i = 1, \ldots, N \tag{16.12}$$

and

$$X_i^{\text{comp}}\left(\vec{\mathbf{p}} + \vec{\mathbf{t}}; \overline{U}^0\right) = \pi_i\left(\vec{\mathbf{p}}\right) \quad i = 2, \ldots, N \tag{16.13}$$

produce different vectors of producer prices $\vec{\mathbf{p}}$ for any given $\vec{\mathbf{t}}$. Moreover, the compensated $\vec{\mathbf{p}}$ depends as well on the good picked for compensation (good 1 in all of our examples). Notice, too, that the level of pure profits also differs in the two equilibria, equal to $\pi(\vec{\mathbf{p}}_A)$ in the actual equilibrium and $\pi(\vec{\mathbf{p}}_c)$ at the compensated equilibrium. (So will the tax revenues collected, but this is true even for the linear technology case.) For the loss measure to be well defined, then, compensation in some stated good must actually be paid by some agent outside the economy so that the compensated price vectors are observed. Without actual compensation, it is not clear how to evaluate loss. In particular, evaluating loss using Eqns (16.3) and (16.9) at actual tax and price vectors $(\vec{\mathbf{t}}, \vec{\mathbf{p}}_A, \vec{\mathbf{q}}_A)$, the only vectors actually observed, is not a well-defined theoretical measure.

7. The market clearance equation, Eqn (16.6), are also required with general technology.

Hicks' Compensating Variation versus Hicks' Equivalent Variation Welfare Measures

A final comment concerns the choice of compensation. Our analysis has made use of Hicks' Compensating Variation (HCV), defining the compensated equilibrium at the new prices and the original, before-tax utility level. Most incidence studies are presented in the spirit of Hicks' Equivalent Variation (HEV), defining the compensated equilibrium at the original before-tax prices and the new, with-tax utility level. The justification for using the HEV is that it is a money index of utility, since compensation is always measured at the same set of relative prices.

The problems in mixing actual and compensated equilibria discussed above still apply to the HEV framework, however. The expenditure function measures the income the consumer would be willing to sacrifice to return to the original before-tax prices. Placing the consumer on the actual new after-tax indifference curve at the tangency of the original, before-tax prices would bring the economy to a different point on the production−possibilities frontier than the actual with-tax equilibrium, with a different vector of producer prices under general technology. Therefore, if taxes were levied and the consumer paid compensation out of lump-sum income to remain on the new actual after-tax utility level at the before-tax prices, the economy would reach a compensated general equilibrium with vectors of consumer and producer prices different from either the actual before-tax or actual after-tax price vectors. Equation (16.13) would be needed to solve for the compensated price vectors given t, with actual after-tax U instead of U^0, and the discussion surrounding Eqn (16.13) applies. Mixing actual and compensated equilibria is still not legitimate, although it is commonly done in the incidence literature.

The Relative Price Change Measure of Incidence

Because of the difficulties in mixing actual and compensated equilibria, many researchers have been content to compute actual changes in consumer prices as *the* measure of incidence, stopping short of relating the price changes directly to changes in welfare in any formal manner. This is obviously a pragmatic compromise. The resulting measures have no particular theoretical justification, but at least they can be computed fairly easily from the actual data.

Using the profit function to represent production, the procedure for computing price changes for differential incidence can be represented as a three-step process, already outlined earlier. First, totally differentiate the actual government budget constraint with respect to the tax rates being changed (usually two of them) to determine the exact changes required to hold the budget surplus constant; for example:

$$\mathrm{d}S = 0 = \frac{\partial(tx^A + pZ)\mathrm{d}t}{\partial t} \quad \left(\text{in vector notation}\right) \tag{16.14}$$

Second, totally differentiate the actual market clearance relationships

$$X_i^{\text{actual}}\left(\vec{\mathbf{p}}_A + \vec{\mathbf{t}}; I\right) = \pi_i\left(\vec{\mathbf{p}}_A\right) + Z_i \quad I = 1, \ldots, N \tag{16.15}$$

with respect to the tax rates to solve for the producer price changes given the changes in taxes determined from differentiating the government budget constraint. (The demand curve should have an income term to allow for the possibility of deadweight loss that reduces real income, even if there is zero pure profit or loss from production. If pure profits exist, it is natural to assume they are received by the consumer as income.)

Finally, use the relationship

$$\mathrm{d}q_i = \mathrm{d}p_i + \mathrm{d}t_i \quad i = 1, \ldots, N \tag{16.16}$$

to determine the resulting changes in consumer prices.

The price changes could be directly related to welfare losses by positing an indirect utility function of the form $V(\vec{\mathbf{q}})$ and totally differentiating it to obtain (in vector notation):

$$\mathrm{d}V = -\lambda X \mathrm{d}q, \tag{16.17}$$

where $\lambda =$ the marginal utility of income. $\mathrm{d}V/\lambda$ represents a money index of utility, but it is path dependent and therefore not uniquely valued for nonmarginal price changes, in general. Consequently, incidence analysis using the change-in-relative-prices measure often concludes the formal analysis with the price changes. The link to consumer's welfare is then simply presented in heuristic terms, in the form of general statements about who gains and who loses (with many consumers).

In summary, then, the theory of tax incidence presents a quandary even for simple one-consumer-equivalent economies. Despite the obvious motivation for developing empirical measures of tax incidence, there appear to be no obvious candidates for the task unless production technology is linear. With general technologies, unambiguous measures of welfare loss involve compensated equilibria that cannot be observed in practice, and observed tax and price vectors offer, at best, only intuitive guidance to welfare losses. As a practical matter, economists may have to be content with measures of price changes in response to different sets of taxes that leave the government budget surplus unchanged, especially given that production technologies are general and not linear.

The only firm conclusion one can draw is that if the incidence of a given set of taxes is to have any meaning in

a general equilibrium context, tax incidence must be defined in such a way to render the impact of a tax only indirectly relevant to the incidence measure. Tax revenues (or the resulting budget surplus) from distorting taxes must be returned lump sum to the consumer to have a well-defined problem focusing on a single tax. The actual tax payment can affect incidence only through its influence on the amount that market prices change in response to the tax. Regardless of whether one chooses the income compensation or change-in-relative-price approach, the final incidence measure is fully determined by the resulting changes in the general equilibrium price vectors.

THE EQUIVALENCE OF GENERAL TAXES

Although the income compensation and change-in-actual-price measures of incidence approach the problem from different perspectives, they each imply the following important result: In a perfectly competitive, profitless economy, in which tax revenues (or the budget surplus) are always returned lump sum, any two sets of taxes have identical incidence if they generate the same changes in relative prices.

Consider, first, the relative price measure of incidence. If production is profitless and tax revenues are returned to the consumers, actual consumer demands (factors suppliers) are functions of only relative prices. Producers' supply (input demand) relationships are also functions of only relative prices. Therefore, two sets of taxes that generate the same vector of relative prices generate the same with-tax general equilibrium. Consequently, they must have the same incidence by the relative price criterion.

That two sets of taxes generating the same vector of relative prices have the same incidence using the income compensation measure can be most easily demonstrated as follows. Suppose compensation is paid in good 1 so that the market for good 1 remains uncleared in the compensated equilibrium. With compensation defined in terms of good 1, loss can be represented as

$$L(t) = M_1(\vec{\mathbf{q}}; \overline{U}) - \pi_1(\vec{\mathbf{p}}) \tag{16.18}$$

Equation (16.18) measures the difference between the amount of good 1 required for compensation and the amount of good 1 available to the consumer from production, given that all other compensated demands (factor supplies) have been satisfied $(M_i(\vec{\mathbf{q}}; \overline{U}) = \pi_i(\vec{\mathbf{p}}),$ $i = 2, \ldots, N)$. But both M_1 and π_1 are homogenous of degree zero in prices. Therefore, any two taxes creating the same vector of relative prices must generate the same deadweight loss or incidence.[8]

These considerations lead to a well-known theorem on the equivalence of general taxes that applies to a competitive, profitless economy. A general tax has the following properties: (1) if levied on a single consumer good (factor supply), all consumers pay the same tax rate; (2) if levied on more than one good (and/or factor), property (1) holds for each taxed good (factor), and all the taxed goods (factors) are taxed at the same rate.

Theorem: The Equivalence of General Taxes

Let $(X_1, \ldots, X_N)$ be the vector of goods and factors for a competitive, profitless economy with producer prices $(p_1, \ldots, p_N)$. Levy a general ad valorem tax at rate t, paid by consumers, on any subset of the goods and factors, say $X_1, \ldots, X_k$, such that the consumer prices are $q_i = P_i(1 + t)$, $i = 1, \ldots, K$, and $q_j = p_j, j = k + 1, \ldots, N$. It is always possible to replace the tax with another general ad valorem tax at rate t^* on the remaining goods and factors $(X_{k+1}, \ldots, X_N)$ such that the two taxes have the same incidence.

Notice that if $(X_1, \ldots, X_k)$ is the subset of goods and $(X_{k+1}, \ldots, X_N)$ is the subset of factors, the theorem establishes the equivalence between a general sales tax and a general income tax (or general value-added tax). Dividing the goods and factors in this way is not necessary; however, any two-way division will do. For example, the theorem also establishes the equivalence between a tax on any one good (or factor) and a tax on all the remaining goods and factors.

Proof: With the ad valorem tax t on the subset $(X_1, \ldots, X_k)$, the following relationships hold in equilibrium:

$$\mathrm{MRS}_{ij} = \frac{p_i}{p_j} = \mathrm{MRT}_{ij} \quad i,j \text{ both in} (k+1, \ldots, N) \tag{16.19a}$$

$$\mathrm{MRS}_{ij} = \frac{p_i(1+t)}{p_j(1+t)} = \frac{p_i}{p_j} = \mathrm{MRT}_{ij} \quad i,j \text{ both in} (1, \ldots, k) \tag{16.19b}$$

$$\mathrm{MRS}_{ij} = \frac{p_i(1+t)}{p_j} = (1+t)\mathrm{MRT}_{ij} \quad i \text{ in} (1, \ldots, k) \quad j \text{ in} (k+1, \ldots, N) \tag{16.19c}$$

With the ad valorem tax t^* on the subset $(X_{k+1}, \ldots, X_N)$, the following relationships hold:

$$\mathrm{MRS}_{ij} = \frac{p_i}{p_j} = \mathrm{MRT}_{ij} \quad i,j \text{ both in} (1, \ldots, k) \tag{16.19a$'$}$$

$$\mathrm{MRS}_{ij} = \frac{p_i(1+t^*)}{p_j(1+t^*)} = \frac{p_i}{p_j} = \mathrm{MRT}_{ij} \quad i,j \text{ both in} (k+1, \ldots, N) \tag{16.19b$'$}$$

8. The government's purchase and sale of Z_i can be included in Eqn (16.18) and the market clearance equations without affecting the result. The Z_i are under the control of the government and therefore exogenous. Also, they do not alter the homogeneity properties of the M_i and π_i.

$$\mathrm{MRS}_{ij} = \frac{p_i}{p_j(1+t^*)} = \frac{1}{(1+t^*)}\mathrm{MRT}_{ij} \quad i \text{ in}(1,\ldots,k) \quad j \text{ in}(k+1,\ldots,N) \tag{16.19c'}$$

In a profitless economy, only relative prices matter in determining the general equilibrium. For the taxes to be equivalent, then, Eqn (16.19c) must equal Eqn (16.19c′), which requires that t^* be set such that

$$\frac{p_i(1+t)}{p_j} = \frac{p_i}{p_j(1+t^*)} \tag{16.20}$$

or

$$(1+t)(1+t^*) = 1 \tag{16.21}$$

with t and t^* defined as decimal fractions.

Implications

1. If $t > 0$, then $t^* < 0$. For example, if $t = 100\%$ ($t = 1$), t^* must be set equal to -50% ($t^* = -1/2$). Thus, a general sales tax of 100% on all goods is equivalent to a 50% tax on all factors (factors are measured negatively, so that a negative t^* applied to a factor supply is a tax). If the subsets $X = 1, \ldots, k$ and $i = k+1, \ldots, N$ each include a mix of goods and factors, then some elements of *each* subset are taxed and others subsidized, depending on whether they are goods or factors.
2. The numerical example illustrates that one of the ad valorem rates, in the case t^*, is applied to the gross-of-tax price and the other, in this case t, to the net-of-tax price. This merely reflects the fact that the producer price for factors is a gross-of-tax price, while the producer price for goods is a net-of-tax price. To see this, consider Fig. 16.3 (goods market) and Fig. 16.4 (factor market) and assume that the economy consists of a single good and a single factor. An ad valorem tax paid by the consumer shifts the demand curve down in the goods market and the supply curve up in the factor market. For q_i/q_2 to be the same for either tax, the tax rate applied to p_1 must exceed the tax rate applied to p_2 in absolute value. In the graphical example, t raises q_1 100% above p_1, the net-of-tax price, and t^* lowers q_2 50% below p_2, the gross-of-tax price. At these rates,

$$\frac{q_1}{q_2} = \frac{p_1(1+1)}{p_2} = \frac{p_1}{\left(1-\frac{1}{2}\right)p_2} = 2\left(\frac{p_1}{p_2}\right) \tag{16.22}$$

3. In these examples, the consumers actually pay the tax. The same theorem applies if the producers paid the tax, the only difference being that

$$p_i = (1+t)\cdot q_i \quad i = 1,\ldots,k \tag{16.23}$$

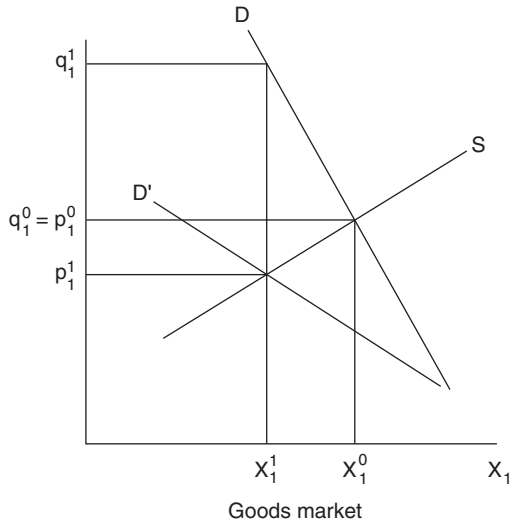

FIGURE 16.3

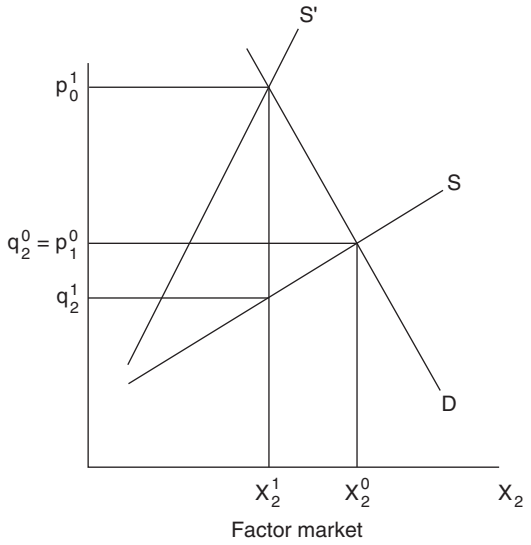

FIGURE 16.4

$$p_i = (1+t^*)\cdot q_i \quad i = k+1,\ldots,N \tag{16.24}$$

$(1 + t)\,(1 + t^*) = 1$ is still required for equivalence. In Figs 16.3 and 16.4, the opposite curves would shift, with t applied to the gross-of-tax price in the goods market and t^* to the net-of-tax price in the factor market. Whether the impact of a general tax falls on the buyer or supplier in any market can never affect its incidence. This is a particular instance of the general principle that it does not matter which side of a market is taxed. Any tax levied on consumers can in principle be duplicated by a tax levied on producers. In practice, however, the government may prefer to tax one side or the other. For instance, an income tax can more easily take into account the personal characteristics of families and individuals than can a sales tax.

4. The two taxes generate the same tax revenue in either the compensated or actual general equilibria. Consider, first, the compensated equilibrium. By design, the two taxes generate the same relative prices, the same compensated equilibria, and the same deadweight loss. Thus,

$$L(t) = M\left(\overrightarrow{\mathbf{q}}^{t}; \overline{U}^{0}\right) - \sum_{i=1}^{k} t p_i X_i^{\text{comp}} - \pi(\overrightarrow{\mathbf{p}}^{t}) \qquad (16.25)$$

$$L(t^*) = M\left(\overrightarrow{\mathbf{q}}^{t}; \overline{U}^{0}\right) - \sum_{i=k+1}^{N} t^* p_i X_i^{\text{comp}} - \pi(\overrightarrow{\mathbf{p}}^{t^*}) \qquad (16.26)$$

and $M(\overrightarrow{\mathbf{q}}^{t}; \overline{U}^{0}) - \pi(\overrightarrow{\mathbf{p}}^{t}) = M(\overrightarrow{\mathbf{q}}^{t^*}; \overline{U}^{0}) - \pi(\overrightarrow{\mathbf{p}}^{t^*})$. Therefore,

$$\sum_{i=1}^{k} t \cdot p_i X_i^{\text{comp}} = \sum_{i=k+1}^{N} t^* p_i X_i^{\text{comp}} \qquad (16.27)$$

Turning to the actual equilibrium, the tax revenues with each tax are

$$T_t^A = \sum_{i=1}^{k} t \cdot p_i^A X_i^A = t \cdot \sum_{i=1}^{k} p_i^A X_i^A \qquad (16.28)$$

$$T_{t^*}^A = \sum_{i=k+1}^{N} t^* p_i^A X_i^A = t^* \cdot \sum_{i=k+1}^{N} p_i^A X_i^A \qquad (16.29)$$

With perfect competition and zero profits, the actual goods demands and factor supplies are functions of only relative prices. Therefore, t and t^* generate the same general equilibria with the same producer prices. Also, $\sum_{i=1}^{N} p_i Y_i = 0$, where Y_i is the supply of (demand for) good (factor) i. Therefore, from market clearance, $\sum_{i=1}^{N} p_i X_i = 0$,

$$\sum_{i=1}^{k} p_i X_i = -\sum_{i=k+1}^{N} p_i X_i \qquad (16.30)$$

Given the design of the taxes, under t, $\sum_{i=k+1}^{N} q_i^t X_i = \sum_{i=k+1}^{N} p_i X_i$; under t^*, $\sum_{i=1}^{k} q_i^{t^*} X_i = \sum_{i=1}^{k} p_i X_i$. The value of the nontaxed goods to the consumer is the same under both taxes (except for sign). But the X_i includes all the goods and factors, and there is no lump-sum income. Therefore, the consumer's budget constraint is $\sum_{i=1}^{N} q_i X_i = 0$ under t and t^*, which implies that the value to the consumer of the taxed goods and factors is also equal under the two taxes (except for sign): $\sum_{i=1}^{k} q_i^t X_i = -\sum_{i=k+1}^{N} q_i^{t^*} X_i$. Finally, $q^t = (1 + t)p_i$, $i = 1$ to k, and $q^{t^*} = (1 + t^*)P_i$, $i = k + 1$ to N. Therefore, subtracting Eqn (16.30) from the value of the taxed goods yields

$$\sum_{i=1}^{k} t p_i X_i = \sum_{i=k+1}^{k} t^* p_i X_i \qquad (16.31)$$

The tax revenues are equal under the two taxes (recall that t and t^* have opposite signs).

The only difference between the two taxes is the absolute value of the consumer prices. For example, comparing a sales tax with an income tax, the goods and factor prices to the consumer are *both* higher under the sales tax. The different absolute prices have no effect on the consumer's welfare or on the tax revenues collected, however.

5. The theorem applies only to a one-period, profitless economy. If there were pure profits or losses in production, they would have to be accounted for to define incidence equivalence. And designing equivalent taxes in a multiperiod model is much more difficult, since all contemporaneous and intertemporal price ratios have to be equal with the two sets of taxes. For example, we saw in Chapter 11 that interest income would have to be deductible under an income tax to make it equivalent to an expenditures (sales) tax. Despite these qualifications, the theorem on the equivalence of any two general taxes that span the entire set of goods and factors is one of the more powerful results in all of tax incidence theory, especially since it applies for either measure of tax incidence.

MEASURING TAX INCIDENCE: A MANY-CONSUMER ECONOMY

Our discussion of tax incidence measures and methodology in a one-consumer economy is all preliminary. Tax incidence theory is ultimately concerned with the relative burdens from taxation suffered by various consumers or groups of consumers within the economy. As such, it requires analysis within the context of a many-person consumer economy.

Unfortunately, it is not entirely clear how to conceptualize a valid incidence analysis for the many-person economy. There is, at the outset, a fundamental and ambiguous issue centered around the question of point of view: What matters to incidence theory in a many-person economy—the losses suffered by each of the (H) consumers in the economy as *individually* perceived or the *aggregate* loss from a social welfare perspective? Optimal second-best policy analysis certainly requires the aggregate viewpoint, as was demonstrated for the many-person optimal tax problem in Chapter 14. Nonetheless, one could reasonably argue that incidence analysis merely tries to describe the pattern of burdens as perceived by each individual (group of) consumer(s). This view, in effect, says

that tax incidence is meant to fall within the domain of the positive theory of the public sector, not the normative theory.

The Individual Perspective on Incidence

Although the individual perspective on tax incidence is certainly appealing, how to maintain an individual perspective in a many-person economy is not at all clear. The same issue arose in Chapter 14 when we discussed deadweight loss in the context of a many-person economy, because incidence and loss are equivalent under a theoretically appropriate measure of the incidence of distorting taxes in the one-consumer case. Consider, for example, the problem of measuring the incidence of a single tax from the viewpoint of each individual's loss. Determining the incidence of a single tax requires, at the outset, a specific assumption about how the revenue is given back to the consumers. The natural assumption for incidence analysis is that each consumer receives lump sum, exactly the revenue he or she pays. Any other distribution of the revenues blurs the focus on the incidence of the tax in and of itself.

There are three possible ways to view each individual's loss with this assumption, two of them virtually identical with the one-consumer case: the individuals' deadweight losses, the change in relative prices, and any one person's deadweight loss.

Individual Deadweight Loss

One possibility is to compute the loss function for each individual as

$$L^h(\vec{\mathbf{t}}) = M^h\left(\vec{\mathbf{q}}; \overline{U}^0\right) - \sum_{i=1}^{N} t_i X_{hi}^{\text{comp}} - \pi^h(\vec{\mathbf{p}}) \quad (16.32)$$

using the HCV measure, where $\pi^h(\vec{\mathbf{p}})$ = person h's share of pure profits (losses) from private production and $(\vec{\mathbf{q}}, \vec{\mathbf{p}})$ are the consumer and producer price vectors at the compensated equilibrium. Then compare individual losses. Although this would give unambiguous individual measures of loss that could be compared across consumers, it suffers the same defects as its one-consumer counterpart, with one additional problem. As already noted in the discussion of one-person measures, it implies a conceptual experiment in which not only all the tax revenues are returned lump sum to each individual exactly as collected, but also one in which each person simultaneously receives additional lump-sum income (from an agent outside the economy) to fully compensate him for the given pattern of tax rates. Moreover, the compensated equilibrium would not exhibit the same vector of market prices as the actual general equilibrium under general technology unless the compensation actually takes place. It suffers the further handicap in a many-person context that the government would generally not be interested in compensating individuals in this way even if it could, since compensating each individual fully for his or her self-perceived loss requires more resources than are needed to restore the original level of social welfare. This last point was discussed in Chapter 4.

Change in Relative Prices

The second option is to compute the actual change in market prices $(\vec{\mathbf{q}}_A, \vec{\mathbf{p}}_A)$ and infer the pattern of burdens from these changes; although, how such inferences are to be made is difficult to see. As noted in the one-consumer case, the government could use each individual's indirect utility function to compute individual money indexes of utility loss, although the value of such indexes is questionable.[9]

One Person's Deadweight Loss

A third option, not open in the one-consumer case, is to focus on the welfare loss of a single person (or one "small" group of consumers) and ask how much income this person (group) would require as compensation for the actual change in market prices resulting from the tax. That is, compute for some person, *but only for that person,*

$$L^h(\vec{\mathbf{t}}) = M^h\left(\vec{\mathbf{q}}_A; \overline{U}^0\right) - \sum_{i=1}^{N} t_i X_{hi}^{\text{comp}}\left(\vec{\mathbf{q}}_A; \overline{U}^0\right) - \pi^h(\vec{\mathbf{p}}_A) \quad (16.33)$$

where the loss function is evaluated at the actual with-tax market prices. Since only one person (or one "small" group) is conceptually being compensated, this compensation would presumably leave the actual market prices unchanged, so that this conceptual experiment is well defined. Thus, we can consider the loss suffered by one person (group) as he (it) perceives the loss; although, we cannot do this for all people (groups) and compare results. Compensating everyone simultaneously at the actual market prices is *not* a well-defined conceptual experiment with general technologies. This could only be done unambiguously if technology were linear, in which case the observed vector of consumer prices obtains no matter how a conceptual compensation experiment is defined.

The same three options apply to differential incidence, in which one tax is substituted for another. Presumably one would be interested in computing the tax changes necessary for a constant government budget surplus at the actual equilibrium. Given these tax changes, the question remains whether compensation tests could be mixed with actual

9. Peter Diamond takes this approach in Diamond (1978).

market results, a question that has been fully discussed in the context of a one-consumer (equivalent) economy.

The Aggregate Social Welfare Perspective on Incidence

There is no ambiguity if the aggregate social welfare point of view is adopted. One would then compute changes in actual market equilibria and their resulting effects on the social welfare function. The aggregate differential incidence problem has already been presented at the end of Chapter 14 for a profitless, competitive economy with no government production. Recall that there are two key relationships. One is the government budget constraint:

$$\sum_h \sum_i t_i X_{hi} = \overline{T} \tag{16.34}$$

which can be totally differentiated to determine the changes in tax rates necessary to hold tax revenues constant.[10] The other is Eqn (14.88) (in vector notation):

$$dW = \left| \left[-(1-\beta)'X \right] - t' \frac{\partial X}{\partial q} \right| E^{-1} \frac{\partial Y}{\partial p} dt \tag{16.35}$$

which relates changes in social welfare to changes in tax rates.

$\beta = \begin{bmatrix} \beta^{l} \\ \beta^{h} \\ \beta^{H} \end{bmatrix}$ is the vector of social marginal utilities of income.

The aggregate perspective thus formulates the differential incidence question as determining which of two sets of taxes generates the higher level of social welfare. Although this is certainly a well-defined general equilibrium problem, economists have typically adopted an individual perspective when analyzing the incidence of taxes.

THE HARBERGER ANALYSIS

Arnold Harberger's 1962 analysis of the incidence of corporate income tax stands as a landmark without rival in the literature on tax incidence theory. Its contributions were twofold. In the first place, his study firmly established the fundamental principle that incidence analysis, properly conceived, requires a full general equilibrium model of the underlying economy. Second, Harberger developed the methodology for measuring incidence in terms of changes in actual general equilibrium consumer and producer prices, focusing primarily on changes in factor prices. Although, as noted above, this measure cannot possibly be the definitive measure of tax incidence, no other single measure is infallible either. Many tax theorists have chosen Harberger's method of analysis in their own studies, regardless of the tax being analyzed. They do so because Harberger's model gives a good intuitive sense of how the market economy spreads the burden of a tax beyond its point of impact in determining the incidence of the tax. For all these reasons, Harberger's study of the corporate income tax deserves careful attention. It also happens to be, somewhat ironically, an excellent vehicle for demonstrating the limitations of the change-in-actual-prices measure of incidence as a measure of true economic burdens. Thus, it serves as an appropriate conclusion to the chapter.

For his analytical framework, Harberger chose a one-consumer (equivalent), profitless, perfectly competitive-market economy with general, constant returns to scale (CRS) production technology. His basic methodology can be stated very simply in terms already outlined in the preceding sections of this chapter. First, he chose to analyze the incidence of a single "small" tax in which the revenues were returned to the consumer lump sum. Specifically, Harberger posited a single tax on the use of capital services by all firms in one of two sectors within the economy, the corporate sector,[11] the proceeds of which are spent by the government exactly as the consumer would have spent them. This assumption is equivalent to returning the taxes lump sum, and it automatically maintains budgetary balance (at level zero) and consumers' lump-sum income (also at zero with CRS).[12] Once the tax rate is specified, all that is required to determine the resulting price changes is differentiating the market clearance equations of the form:

$$D^i[p(t) + t] = \pi_i[p(t)] \quad i = 1, \ldots, N \tag{16.36}$$

where

$D^i(\ )$ = demand (supply) for good (factor) i by the consumer.
π_i = the supply (demand) of good (factor) i, the first derivative of the competitive profit function for the economy.

These equations incorporate all the relevant information on preferences (through the demand relations), production technologies (through the profit function), and market clearance, the three elements needed to determine the full general equilibrium. They solve for the producer price changes, after which the consumer price changes follow directly from the price relationships, $q_i = p_i + t_i$, for all $i = 1, \ldots, N$.

10. With government production, the overall surplus should be held constant.

11. Notice that this is a "specific" or "selective" tax as opposed to a "general" tax, since only a subset of all the demanders of capital is taxed.

12. For a more careful discussion of the effect of this tax and transfer on the consumer's income, see page 563.

With N goods and factors, it is impossible to determine a priori how these prices will change, in general. In much simpler economies, however, the pattern of price changes is often predictable. Harberger chose the standard two-good, two-factor model used in most geometric presentations of general equilibrium analysis, and was able to describe precisely how the various demand and production parameters of this model determine the changes in the wage-rental ratio resulting from the corporate tax. He concluded that the capitalists would bear all or nearly all of the tax burden under most reasonable values of these parameters.

Harberger's analytics are much more complicated than solving Eqn (16.36). He works directly with the underlying production functions for the economy, rather than the profit functions, in order to highlight the manner in which production parameters influence the pattern of tax incidence. Most other researchers have followed his lead in this regard. Consequently, Harberger's general equilibrium model contains five basic sets of assumptions:

1. There are two goods, X and Y, each produced by two factors of production, capital (K) and labor (L). Consumers supply the factors in absolutely fixed amounts, a standard assumption that permits one to draw the pareto-optimal production frontier in capital—labor space, because the boundaries of the (K,L)-Edgeworth box are fixed.
2. Production is CRS for each good, according to the production relationships:

$$X = X(K_X, L_X) \tag{16.37}$$

$$Y = Y(K_Y, L_Y) \tag{16.38}$$

   Furthermore, the two industries are unequally factor intensive, meaning that $K_X/L_X > K_Y/L_Y$ (X being relatively capital intensive) *or* $K_X/L_X < K_Y/L_Y$ (Y being relatively capital intensive) at any given feasible factor price ratio, P_K/P_L. This assumption, along with CRS, generates a production-possibility frontier that is uniformly concave to the origin, so that general equilibrium price ratios vary systematically as the economy moves along the frontier. The CRS assumption rules out the possibility of pure profits or losses at any competitive equilibrium.
3. The model is static; there is no saving in the economy, even though capital is one of the factors of production. Also, all markets are competitive, an assumption that[13] has two very important implications:
   a. The equilibrium is characterized by full employment of all resources so that $K_X + K_Y = \overline{K}$ and $L_X + L_Y = \overline{L}$, where $\overline{K}$ and $\overline{L}$ are the fixed factor supplies.
   b. In equilibrium, all consumers pay the same prices for X and Y no matter where purchased, and they must receive the same returns for their factors of production whether they are supplied to industry X or industry Y. Also, the equilibrium factor prices equal the value of their marginal products in each industry.
4. The government levies a "small" tax on the use of capital services in industry X, identified as the corporate sector. There are no other taxes in the economy. To dispose of the revenue, the government spends the proceeds exactly as the consumers would have had they kept the revenue but were confronted with the new general equilibrium vector of prices. As mentioned, this is equivalent to returning the tax revenues lump sum. It also preserves the total level of national income within the economy.
5. Since Harberger does not introduce a social welfare function, he implicitly assumes a one-consumer equivalent economy.

Geometric-Intuitive Analysis

With these five sets of conditions, Harberger is able to describe the change in the factor price ratio P_K/P_L in response to the tax. The changes in factor incomes accruing to capital and labor as a result of the changes in P_K/P_L measure, for Harberger, the true economic burdens of the tax borne by capital and labor. Before turning to his analytical equations, which are fairly complex, let us first develop a feel for Harberger's results by undertaking a geometric-intuitive analysis of the general equilibrium response to the tax in a simple two-good, two-factor economy.

A tax on the use of capital in industry X has the immediate effect of driving a wedge between the returns to capital in the two sectors. Investors in industry X receive the net-of-tax return $(P_K^0 - T_{KX})$, where T_{KX} is the unit tax on capital in industry X. Investors in industry Y continue to receive the gross-of-tax return, P_K^0. Presumably firms in industry X try to increase P_X by an amount sufficient to restore the original rate of return P_K^0. Whether or not they succeed depends upon the demand elasticity for good X. In a two-good economy, one would expect the demand for X to have some price elasticity and that X and Y would be substitutes. Therefore, the demand for Y could increase in response to a rise in the price of X. If this is true, then P_x does not rise sufficiently to cover the tax in the short run, generating losses in industry X. At the same time, profits arise in industry Y, and firms have an incentive to shift resources from X to Y in order to equalize the returns to capital in both industries.

13. Harberger relaxes this assumption in the last part of his article by permitting monopoly power in the market for X, the taxed corporate sector.

What happens then depends on relative factor intensities. Suppose X is relatively capital intensive. If so, then at the initial factor price ratio P_K^0/P_L^0, industry X is releasing capital and labor in different proportions from those desired by industry Y, generating excess supply in the capital market and excess demand in the labor market. The factor price ratio P_K/P_L begins to fall, and *both* industries respond by becoming more capital intensive as factor markets continue to equate factor prices with values of marginal products. Equilibrium is achieved only when full employment is restored in *both* factor markets. The amount of factor price change required to bring this about depends not only on the relative factor intensities but also on elasticities of substitution between capital and labor in both industries.

To give one extreme example indicating how elasticities of substitution matter, if the elasticity of substitution between capital and labor is infinite (straight-line isoquants) in the untaxed sector, then there can be only one equilibrium factor price ratio for industry Y, the original P_K^0/P_L^0. For a given P_L, the demand schedule for capital in industry Y is perfectly elastic at the original P_K^0. Hence, capital shifts until $(P_K - T_{KX})$ in industry X just equals P_K^0 in industry Y, and the return to capital does not fall as a result of a tax on capital in industry X. This case is depicted in Fig. 16.5.

Finally, returning to the goods markets, the shift in resources to industry Y tends to lower the goods price ratio P_X/P_Y. This follows because the price of capital has fallen relative to the price of labor, and industry X is the relatively capital-intensive industry. Consequently, production (marginal) costs should fall in industry X and rise in Y. What this says, in effect, is that the long-run supply curves (marginal costs) for both goods are expected to be upward sloping with CRS and unequal factor intensities. Overall, the final change in the goods price ratio is indeterminant a priori.

Figure 16.6 gives one possible outcome in which there is no change in P_X/P_Y. The shift in the demand curve for Y in response to the original increase in the price of X is just enough to restore (the posited) equality of P_Y and P_X, so that the ratio of these prices remains unchanged. The tax tends to increase P_X relative to P_Y because costs are rising (relatively) in X, but the demand response moves the prices in the other direction. In fact, because Harberger focuses entirely on the changes in the factor price ratio P_k/P_L for his measure of incidence, he is implicitly assuming that there is no change in the equilibrium goods price ratios, exactly as depicted in Fig. 16.6 (or at least that the final change is "small" and can be ignored).

In general, as the economy moves along its productions—possibilities frontier from capital-intensive X to labor-intensive Y, P_K/P_L falls to maintain full employment and P_Y/P_X rises to reflect the relative cost changes in the two industries. The wrinkle with a tax on capital used only in one industry is that it distorts factor markets and drives the economy beneath the production—possibilities frontier. This explains why P_Y/P_X does not have to rise in the new equilibrium in Harberger's analysis.

The descriptive analysis indicates that the incidence of the corporate tax (a tax on the use of capital in sector X) depends on three sets of parameters: the relative factor intensities of the two sectors, the elasticity of substitution between capital and labor in each sector, and the price elasticities of demand for goods X and Y. The analytics uncover a fourth determinant as well, the shares of both capital and labor income originating in each sector. The descriptive analysis also indicates that it is generally not possible to isolate the burden of a tax to one sector of an economy even though the tax is placed selectively within one sector. If investors in the taxed sector suffer a decrease in the return to capital, investors elsewhere suffer the same burden as well, since competitive factor markets equalize returns to capital everywhere in the economy. Furthermore, since markets are interdependent, the tax burdens could spread to other untaxed factors and to consumers through changes in goods prices. In general, then, a selective tax is selective only in its impact, not in its incidence. The market ultimately determines the incidence of a tax, not the legislature.

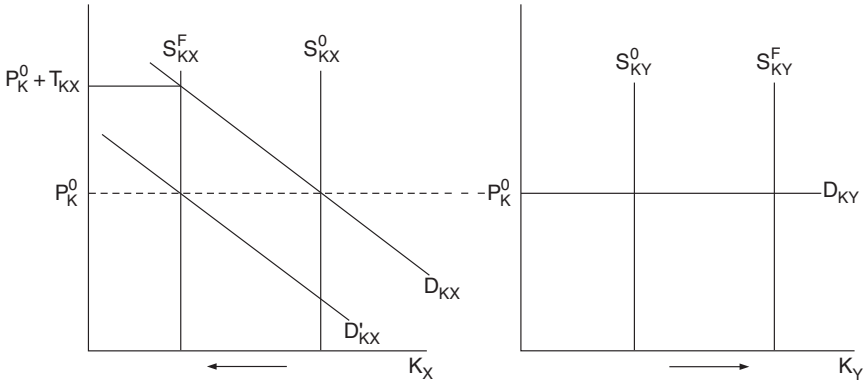

FIGURE 16.5

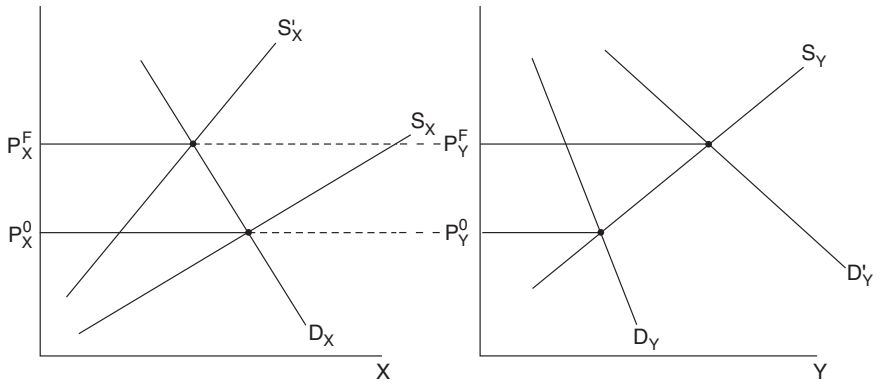

FIGURE 16.6

The Harberger Analytics

Harberger describes the demand, supply, and market clearance equations for his economy with 10 equations designed to highlight changes in the equilibrium values of factor prices and factor supplies in response to the tax on the use of capital in sector X. Since he selects the price of labor as the numeraire, the change in the factor price ratio equals the change in the price of capital, dP_K. The only special feature of his analysis is that all goods and factors are defined in units such that the value of all goods and factor prices in the original pretax equilibrium is one. This is done strictly as a matter of convenience. It implies no loss of generality.

The Demand Equations

Harberger describes the demand side of the general equilibrium model with a single demand equation for X of the form:

$$X = X\left(\frac{P_X}{P_Y}\right) \tag{16.39}$$

With fixed factor supplies, once the change in X is determined in response to the tax, the change in Y follows immediately, since all income is spent on either X or Y. Thus, a separate demand equation for Y is redundant. Also, there is no need to write the demand for X as a function of P_L and P_K. Production is CRS so that factor income exhausts the product—there can be no pure profits or losses. And since the government essentially returns all tax revenue to the consumers, there is no change in the consumers' disposable income even if P_L and P_K change.[14] Finally, given Harberger's one-consumer equivalent assumption, there is only a single demand relationship for X.

14. There may be an income effect in the form of deadweight loss, but Harberger ignores this. We will return to this point later.

Totally differentiating the demand for X yields

$$dX = \frac{\partial X}{\partial\left(\frac{P_X}{P_Y}\right)} \frac{P_Y\, dP_X - P_X dP_Y}{P_Y^2} \tag{16.40}$$

Divide by X to express the change in percentage form:

$$\frac{dX}{X} = \frac{\partial X}{\partial\left(\frac{P_X}{P_Y}\right)} \frac{1}{X}\left[\frac{dP_X}{P_Y} - \frac{P_X}{P_Y^2} dP_Y\right] \tag{16.41}$$

Finally, by multiplying and dividing by P_X/P_Y and with quantity units defined such that $P_X = P_Y = 1$, Eqn (16.41) can be rewritten as

$$\left.\frac{dX}{X}\right|_{\text{demand}} = E\left[dP_X - dP_Y\right] \tag{16.42}$$

where

$$E = \frac{\partial X}{\partial P_X/P_Y} \frac{P_X/P_Y}{X} \tag{16.43}$$

is the demand elasticity for X in terms of the relative prices P_X/P_Y, and dP_X and dP_Y are proportional changes in P_X and P_Y.

The Goods–Supply and Input–Demand Equations

From market clearance, the percentage changes in demand for X must equal the percentage change in supply of X. To determine the percentage change in the supply of X, totally differentiate the production function $X = f(K_X, L_X)$, obtaining

$$dX = \frac{\partial f}{\partial K_X} dK_X \frac{\partial f}{\partial L_X} dL_X \tag{16.44}$$

Therefore,

$$\left.\frac{dX}{X}\right|_{\text{supply}} = \frac{\frac{\partial f}{\partial K_X}}{f} dK_X + \frac{\frac{\partial f}{\partial L_X}}{f} dL_X \tag{16.45}$$

$$\left.\frac{dX}{X}\right|_{\text{supply}} = \frac{\frac{\partial f}{\partial K_X}K_X}{f}\frac{dK_X}{K_X} + \frac{\frac{\partial f}{\partial L_X}L_X}{f}\frac{dL_X}{L_X} \tag{16.46}$$

$$\left.\frac{dX}{X}\right|_{\text{supply}} = \theta_{KX}\frac{dK_X}{K_X} + \theta_{LX}\frac{dL_X}{L_X} \tag{16.47}$$

where

θ_{KX} = the share of capital's income in industry X.
θ_{LX} = the share of labor's income in industry X.

Equation (16.47) follows from Eqn (16.46) because: (1) with CRS, factor payments exhaust the product and (2) factors are paid the value of their marginal products in competitive factor markets. Hence,

$$\frac{\partial f}{\partial K_X}\frac{K_X}{f} = \frac{P_K}{P_X}\frac{K_X}{X} = \theta_{KX}$$

capital's share of income in industry X, and similarly for labor's share. By Walras' law, both the demand and supply equations for Y are redundant.

Turning next to the industries' demands for factors, changes in factor demands can be specified in terms of their direct elasticities of substitution. Define

$$S_Y = \frac{d\log\left(\frac{K_Y}{L_Y}\right)}{d\log\left(\frac{f_{KY}}{f_{LY}}\right)} \tag{16.48}$$

$$S_X = \frac{d\log\left(\frac{K_X}{L_X}\right)}{d\log\left(\frac{f_{KX}}{f_{LX}}\right)} \tag{16.49}$$

where

S_Y = direct elasticity of substitution between capital and labor in industry Y.
S_X = direct elasticity of substitution between capital and labor in industry X.

But, with competitive markets and CRS production,

$$\begin{aligned} d\log\left(\frac{f_{K_X}}{f_{L_X}}\right) &= d\log\left(\frac{f_{K_Y}}{f_{L_Y}}\right) \\ &= d\log\left(\frac{P_K}{P_L}\right) \end{aligned} \tag{16.50}$$

Therefore,

$$d\log\left(\frac{K_Y}{L_Y}\right) = S_Y d\log\left(\frac{P_K}{P_L}\right) \tag{16.51}$$

and

$$d\log\left(\frac{K_X}{L_X}\right) = S_X d\log\left(\frac{P_K}{P_L}\right) \tag{16.52}$$

Consider Eqn (16.51)

$$\begin{aligned} d\log\left(\frac{K_Y}{L_Y}\right) &= \frac{1}{\frac{K_Y}{L_Y}}d\left(\frac{K_Y}{L_Y}\right) = \frac{1}{\frac{K_Y}{L_Y}}\left[\frac{L_Y dK_Y - K_Y dL_Y}{L_Y^2}\right] \\ &= \frac{dK_Y}{K_Y} - \frac{dL_Y}{L_Y} \end{aligned} \tag{16.53}$$

Similarly,

$$d\log\left(\frac{P_K}{P_L}\right) = \frac{dP_K}{P_K} - \frac{dP_L}{P_L} = dP_K - dP_L, \quad \text{with } P_K = P_L = 1 \tag{16.54}$$

Substituting Eqns (16.53) and (16.54) into Eqn (16.51) yields

$$\frac{dK_Y}{K_Y} - \frac{dL_Y}{L_Y} = S_Y(dP_K - dP_L) \tag{16.55}$$

Similarly,[15]

$$\frac{dK_X}{K_X} - \frac{dL_X}{L_X} = S_X(dP_K + T_{KX} - dP_L) \tag{16.56}$$

Market Clearance

Since capital and labor are in fixed supply, capital and labor must move between sectors in equal amounts to maintain full employment. Therefore

$$dK_Y = -dK_X \tag{16.57}$$

$$dL_Y = -dL_X \tag{16.58}$$

Also, the market for X must remain in balance so that

$$\left.\frac{dX}{X}\right|_{\text{demand}} = \left.\frac{dX}{X}\right|_{\text{supply}} \tag{16.59}$$

As indicated earlier, a market clearance equation for Y is redundant given the formulation of the model.

Additional Price Relationships

Because Harberger is interested in changes in relative factor prices as the measure of tax incidence, he presents two additional equations relating changes in the goods prices to changes in the factor prices. Consider, first, the market equilibrium for industry X:

$$P_X X = P_L L_X + (P_K + T_{KX})K_X \tag{16.60}$$

15. Recall that Harberger begins with zero taxes, so that $T_{KX} = dT_{KX}$ and $T_{KX} \approx 0$. Equations (16.47), (16.50), and (16.52) implicitly include T_{KX} in P_K for industry X.

from product exhaustion with CRS. Totally differentiating:

$$P_X dX + X dP_X = P_L dL_X + L_X dP_L + (P_K + T_{KX})dK_X + (dP_K + T_{KX})K_X \tag{16.61}$$

But with competitive pricing,

$$\frac{\partial f}{\partial L_X} = \frac{P_L}{P_X} \text{and} \tag{16.62}$$

$$\frac{\partial f}{\partial K_X} = \frac{P_K + T_{KX}}{P_X} \tag{16.63}$$

Moreover from differentiating the production function:

$$dX = \frac{\partial f}{\partial L_X} dL_X + \frac{\partial f}{\partial K_X} dK_X \tag{16.64}$$

Therefore, from Eqn (16.62):

$$P_X D_X = P_L dL_X + (P_K + T_{KX})dK_X \tag{16.65}$$

and

$$X dP_X = L_X dP_L + K_X(dP_K + T_{KX}) \tag{16.66}$$

Thus,

$$dP_X = \frac{L_X}{X} dP_L + \frac{K_X}{X}(dP_K + T_{KX}) \tag{16.67}$$

With all prices equal to 1, and the level of taxes equal to 0 to a first order of approximation,

$$\frac{P_L}{P_X} = \frac{P_K + T_{KX}}{P_X} = 1 \tag{16.68}$$

so that Eqn (16.67) can be rewritten as

$$dP_X = \theta_{LX} dP_L + \theta_{KX}(dP_K + T_{KX}) \tag{16.69}$$

By similar analysis,

$$dP_Y = \theta_{LY} dP_L + \theta_{KY} dP_K \tag{16.70}$$

Finally, labor is chosen as the numeraire. Thus, $P_L \equiv 1$, and

$$dP_L = 0 \tag{16.71}$$

Summary

Equations (16.42), (16.47), and (16.55)–(16.59), and (16.69)–(16.71) describe the comparative static changes in the general equilibrium quantities and prices. Plugging Eqns (16.69) and (16.70) into Eqn (16.42) and employing Eqns (16.57), (16.58) and (16.59), and (16.71), the ten-equation system can be collapsed into the following three-equation system, with dP_K, dL_X/L_X, and dK_X/K_X as the dependent variables:

$$E(\theta_{KY} - \theta_{KK})dP_K + \theta_{LX}\frac{dL_X}{L_X} + \theta_{KK}\frac{dK_X}{K_X} = E\theta_{KX}T_{KX} \tag{16.72}$$

$$S_Y dP_K - \frac{L_X}{L_X}\frac{dL_X}{L_X} + \frac{K_X}{K_Y}\frac{dK_X}{K_X} = 0 \tag{16.73}$$

$$-S_X dP_K - \frac{dL_X}{L_X} + \frac{dK_X}{K_X} = S_X T_{KX} \tag{16.74}$$

For purposes of tax incidence, the variable of interest is dP_K (=$d(P_K/P_L)$, with $P_K = P_L = 1$ and $dP_L = 0$). Using Cramer's rule and combining terms:

$$dP_K = \frac{\left[E\theta_{KX}\left(\frac{K_X}{K_Y} - \frac{L_X}{L_Y}\right) + S_X\left(\frac{\theta_{LX}K_X}{K_Y} + \frac{\theta_{KX}L_X}{L_Y}\right)\right]T_{KX}}{E(\theta_{KY} - \theta_{KX})\left(\frac{K_X}{K_Y} - \frac{L_X}{L_Y}\right) - S_Y - S_X\left(\frac{\theta_{LX}K_X}{K_Y} + \frac{\theta_{KX}L_X}{L_Y}\right)} \tag{16.75}$$

All the relevant information necessary to determine the incidence of the corporate tax is contained in Eqn (16.75).

Comments on the Solution

1. As indicated in the preliminary intuitive analysis, the change in relative factor prices depends upon the demand elasticity for X, the elasticities of substitution between capital and labor in each industry, the relative capital (labor) intensities in the two sectors, and the share of capital and labor income in each sector. Once the change in the price of capital is obtained, it can then be used to compute changes in capital's income relative to changes in national income as a summary measure of incidence, with all changes measured in units of labor, the numeraire. Because the overall supply of capital is fixed, the change in capital's income is simply dP_K. With $P_K = 1$, dP_K also equals the percentage change in income to capital. National income equals the sum of all factor payments, or

$$I = (P_K + T_{KX})K_X + P_K K_Y + P_L L_X + P_L L_Y \tag{16.76}$$

Totally differentiating and recalling that $T_{KX} = dT_{KX}$ (with $T_{KX} = 0$ initially):

$$dI = (dP_K + T_{KX})K_X + dP_K K_Y + dP_L L_X + dP_L L_Y + (P_K + T_{KX})dK_X + P_K dK_Y + P_L dL_X + P_L dL_Y \tag{16.77}$$

But $P_X = P_Y = P_L = P_K = 1$, $dP_L = 0$, $dL_X = -dL_Y$, and $dK_X = -dK_Y$. Hence,

$$dI = T_{KX}K_X + (K_X + K_Y)dP_K \tag{16.78}$$

and

$$\frac{dI}{I} = \frac{T_{KX}K_X + (K_X + K_Y)dP_K}{K_X + K_Y + L_X + L_Y} \tag{16.79}$$

Three cases are of special interest:

a. Suppose, first, that $dP_K = -T_{KX}K_X/(K_X + K_Y)$. This would leave national income unchanged measured in units of labor, whereas capital's share would fall by the entire amount of the tax revenue. In this case, then, capital can be said to bear the entire burden of the tax.

b. Suppose, second, that $dP_K = 0$. Since $dP_L \equiv 0$, the income of both capital and labor would fall in proportion to their initial share in national income. This would imply equal sharing of the tax burden.

c. Finally, suppose the percentage change in the price of capital net of tax (dP_K) just equals the percentage change in national income. This would imply that labor bears the entire burden of the tax. It occurs if

$$dP_K = \frac{dI}{I} = \frac{[T_{KX}K_X + (K_X + K_Y)dP_K]}{L_X + L_Y + K_X + K_Y}, \text{or} \tag{16.80}$$

$$dP_K = \frac{T_{KX}K_X}{L_X + L_Y} \tag{16.81}$$

2. How the burden is shared between capital and labor depends upon the solution to Eqn (16.75), which in turn depends upon the four demand and production parameters embedded in the right-hand side of the equation. Furthermore, the specific cases mentioned above do not place limits on the possible results. Capital could bear a burden greater than its share of the tax revenue ($dP_K < -T_{KX}K_X/(K_X + K_Y)$); similarly, capitalists could actually gain at the expense of labor despite being taxed ($dP_K > T_{KX}K_X/(L_X + L_Y)$). Harberger presents 10 theorems derived from Eqn (16.75), each highlighting how the four supply and demand parameters determine the final incidence of the tax. We will present three of them, indicating how the three special cases mentioned above might occur. The first two theorems were also suggested by the introductory descriptive analysis of the corporate tax.

a. Labor can bear most of the burden of the tax only if the taxed sector is relatively labor intensive.

Proof: For labor to bear most of the burden of the tax, dP_K must be positive. But examination of Eqn (16.75) reveals that this can only occur if industry X is relatively labor intensive. To see this, consider the denominator. The last two terms can be expected to be positive, by inspection. The first term is also generally positive. E can be expected to be negative. $(\theta_{KY} - \theta_{KX})$ and $(K_X/K_Y - L_X/L_Y)$ must have opposite signs, since if capital's share of income is greater in industry Y, $((\theta_{KY} - \theta_{KX}) > 0)$, then industry Y must be relatively capital intensive, or $(K_X/K_Y - L_X/L_Y) < 0$. Thus, the denominator is positive. Turning to the numerator, its second term can be expected to be negative. Hence, for dP_K to be positive, the first term must be positive and greater in absolute value than the second term. Since $E < 0$, and $\theta_{KX} > 0$, this can only occur if $(K_X/K_Y - L_X/L_Y) < 0$, or if the taxed sector, X, is relatively labor intensive.

The exact conditions for which labor bears precisely the full burden of the tax are not easily stated and will not be derived.

b. If the elasticity of substitution between capital and labor in the untaxed industry is infinite, then capital and labor share equally the burden of taxation.

Proof: Equal sharing of the tax burden requires that $dP_K = 0$. But if S_Y, the elasticity of substitution between capital and labor in the untaxed sector, is infinitely large, dP_K must be equal to zero.

c. If both industries are initially equally factor intensive and each has the same elasticity of substitution between capital and labor, then capital bears the full burden of the tax.

Proof: Capital bears the full burden of the tax, if $dP_K = -T_{KX}\, K_X/(K_X + K_Y)$. If both industries are equally intensive, then $K_X/K_Y = L_X/L_Y$, and Eqn (16.75) reduces to

$$\begin{aligned} dP_K &= \frac{S_X\left(\theta_{LX}\frac{K_X}{K_Y} + \theta_{KX}\frac{L_X}{L_Y}\right)T_{KX}}{-S_Y - S_X\left(\theta_{LX}\frac{K_X}{K_Y} + \theta_{KX}\frac{L_X}{L_Y}\right)} \\ &= -\frac{S_X\frac{K_X}{K_Y}\left(\theta_{LX} + \theta_{KX}\right)T_{KX}}{S_Y + S_X\frac{K_X}{K_Y}\left(\theta_{LX} + \theta_{KX}\right)} \end{aligned} \tag{16.82}$$

But, $\theta_{LX} + \theta_{KX} = 1$. Therefore,

$$dP_K = -\frac{S_X\frac{K_X}{K_Y}T_{KX}}{S_Y + S_X\frac{K_X}{K_Y}} \quad \text{or} \tag{16.83}$$

$$dP_K = \frac{-S_X K_X T_{KX}}{S_Y K_Y + S_X K_X} \tag{16.84}$$

If, in addition, $S_X = S_Y$, then

$$dP_K = \frac{-T_{KX}K_X}{K_Y + K_X} \tag{16.85}$$

3. Harberger presents a large number of conditions for which capital bears the full burden of the tax. Suppose that the elasticity of substitution between capital and

labor equals −1, and further that the elasticity of substitution in demand is also −1. These assumptions are often made in empirical research that builds simple models of the economy and simulates them to determine the effects of some public policy, at least for one of the simulations. Capital bears the full burden of the tax under these assumptions. In fact, Harberger shows that capital bears the full burden of the tax if the elasticities of substitution between capital and labor are equal in both industries and equal as well to the elasticity of substitution in demand between the two goods. The proof of this theorem requires extensive manipulation of Eqn (16.75) so we have chosen not to present it, but this is one of the more striking of the 10 Harberger theorems.

In the final section of his paper, Harberger performs a sensitivity analysis on the U.S. economy, computing $dp_K^{\text{U.S.}}$ for what he believes to be a plausible range of estimates for the various elements on the right-hand side of Eqn (16.75). His analysis leads him to the following conclusion (Harberger, 1962).

It is hard to avoid the conclusion that plausible alternative sets of assumptions about the relevant elasticities all yield results in which capital bears very close to 100% of the tax burden. The most plausible assumptions imply that capital bears more than the full burden of the tax.

Harberger also reworks the analysis to include the special taxation of capital gains and the existence of monopoly elements in the corporate sector. Neither of these considerations affects his basic result, that in all likelihood the incidence of the corporate tax in the United States falls substantially upon capital.

4. Harberger's analysis brings into sharp focus the possible differences between tax incidence measured as changes in actual general equilibrium prices and tax incidence measured as changes in welfare or, equivalently, the lump-sum income required to compensate consumers for a given pattern of taxation. A suitable welfare measure would indicate that the tax described by Harberger generates no burden at all, precisely because his tax is an infinitesimally small change from a zero-tax general equilibrium. With the tax revenues (effectively) returned lump sum to the consumer(s), deadweight loss is an appropriate welfare measure, and we saw in Chapter 13 that deadweight loss is zero for a single, infinitesimally small tax. The first marginal distortion is always free.[16]

5. Harberger had to posit a selective tax in order to have an interesting problem given his framework. Had he chosen a general tax on the use of capital in both sectors the tax could not possibly have generated a burden even using the change-in-actual-price measure of incidence. Clearly a tax on the *supply* of a factor in absolutely fixed supply in a one-consumer-equivalent, static economy cannot generate a burden if the revenues are returned. Any tax on a fixed factor is equivalent to a lump-sum tax, and returning all the revenues would return the economy to its original equilibrium. But a tax has the same incidence effects if applied to either side of a market. Hence, a tax, with transfer, on the total demand for capital would also keep the economy at its original equilibrium.

   Figure 16.7 depicts the case of a tax on all capital. The price of capital remains at P_K^0 whether the suppliers are taxed and S remains unshifted or all firms are taxed and demand shifts to D'_K with the shift equal to the full amount of the tax. Assuming the tax revenue is returned in each case, the consumers suffer no loss in income and no deadweight loss.

6. Harberger's assumption that the corporate tax represents an infinitesimal movement away from a world of zero taxes may be an analytical convenience for illustrating his approach to the measurement of tax incidence, but it is certainly an extreme departure from reality. J. Gregory Ballentine and Ibrahim Eris reworked the original Harberger analysis to include an existing corporate tax, while retaining the assumption of zero taxes elsewhere in the economy (Ballentine and Eris, 1975). The assumption of an existing tax changes such calculations as the share of income going to capital in the taxed industry and the change in tax revenues in response to a marginal change in the tax rate. But the major analytical distinction occurs in the equation determining the percentage change in the demand

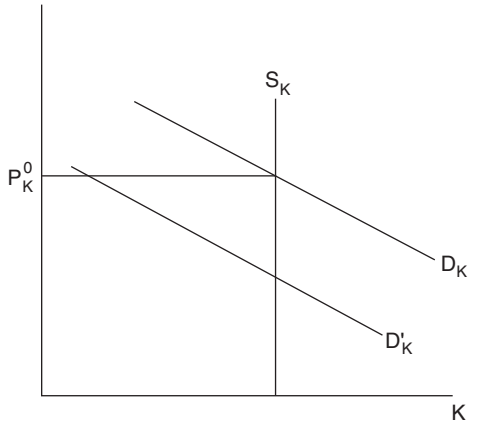

FIGURE 16.7

16. The analysis in Chapter 13 is not strictly relevant since it applies only to general taxes paid by consumers. Harberger considers instead a selective tax paid by some, but not all, of the firms for the use of a specific factor. Nonetheless, it can be easily shown that the deadweight loss is still zero.

X (Eqn (16.42)). Because there is now a marginal deadweight loss from the tax change, the consumers' real income declines, and the change in the demand for X includes this income effect as well as Harberger's relative price effect. When Ballentine and Eris reworked Harberger's empirical sensitivity analysis for the U.S. economy to include the income effect, they found that the burden on capital fell somewhat for plausible values of the income elasticity of demand; although, not enough to alter the conclusion that capital bears the major portion of the tax burden.

7. Harberger's equations are easily modified to analyze per-unit commodity taxes on X or Y or other selective factor taxes such as a tax on the use of capital or labor in Y or a tax on the use of labor in X. For instance, a per-unit commodity tax on X involves adding $T_x(=\mathrm{d}T_x)$ to the right-hand side of Eqn (16.69), with all other equations unchanged (and, of course, removing all T_{KX} terms). A tax on the use of labor in Y requires replacing $\mathrm{d}P_L$ by $(\mathrm{d}P_L + T_{LY})$ in Eqns (16.56) and (16.70), with all other equations unchanged, and similarly for the other taxes.[17] More than one tax change can also be considered with the addition of a government budget constraint, whose derivatives determine the relationship among equal-yield tax alternatives.
8. Finally, Harberger's analysis assumes that the U.S. corporation income tax actually does change the opportunity cost of capital and hence the investment margin within the corporate sector, thereby distorting investors' preferences away from the corporate sector in favor of the unincorporated sector. No attention is given to the characteristics of the tax itself, yet it happens to be a fairly complex tax. For instance, firms are allowed to deduct interest payments on debt and an estimate of depreciation from total returns in computing taxable returns. Moreover net-of-tax returns are taxed again under federal (and state) personal income tax(es), but differentially depending on the exact form of the returns. Dividends and interest income from bonds are taxed as ordinary income, but retained earnings, which ultimately generate capital gains, are taxed at preferential rates, and only when realized. There are also reasonably complex provisions relating to the offset of losses against income. Many other provisions affect the net-of-tax returns as well, too numerous to cite here. The point, however, is that the distortionary effects of this or any other tax depend crucially on its particular design characteristics.

Suppose the corporate tax turned out to be a tax on pure economic profits. In this case, the tax would be lump sum *and* nondistortionary. Corporate investors would simply pay the tax without any adjustment in their investment plans. They would have no incentive to shift resources to the unincorporated sector, and there would be no change in relative prices.

This point has long been understood, yet most economists believe that the corporate income tax is not simply a tax on pure economic profits. Joseph Stiglitz is a notable exception. He argued in a widely cited paper that the tax may well approximate a pure profits tax. This is especially so in a world of certainty, which the Harberger analysis assumes. Given that interest payments on debt and depreciation are both deductible, corporate investment decisions may be independent of the tax.[20]

To see this, consider a firm's decision to borrow \$1 in time $(t-1)$ to finance an additional unit of capital in time t, all other investment plans being unchanged (and optimal). Let

r = the one-period rate of interest on borrowing.
δ = the true rate of economic depreciation.
t_c = the corporate profits tax rate.
$(\partial\pi)/(\partial K_t)$ = the increased operating profits arising from a marginal increase in the capital stock, the gross-of-tax returns to capital.

The decision to invest an additional dollar in time $(t-1)$ leads to $\partial\pi/\partial K_t$ of gross returns in time t, less r dollars of interest costs and 8 dollars of depreciation. If both interest payments and the true economic rate of depreciation are tax deductible, the net-of-tax returns from the investment are $(\partial\pi/\partial K_t - r - \delta)(1 - t_c)$. Hence, the firm should borrow to invest so long as $(\partial\pi/\partial K_t - \delta) \geq r$. Similar analysis for a unit decrease in investment shows that the appropriate disinvestment margin is $(\partial\pi/\partial K_t - \delta) \geq r$. Therefore, the optimal investment plan occurs when $(\partial\pi/\partial K_t - \delta) = r$. The opportunity cost of capital is just r, equal to the gross-of-tax returns net of depreciation; it is independent of t_c. Therefore, if actual depreciation allowances are reasonably close approximations to true economic depreciation, the interest deductibility feature of the corporate tax renders it nondistortionary.[21]

Harberger's analysis turns out to be most compatible with a corporate tax without interest deductibility, in which the net-of-tax returns equal $(1 - t_c)(\partial\pi/\partial K_t - \delta) - r$, or

17. Refer to Mieszkowski (1967), for an analysis of the various possibilities.

20. The seminal paper is Stiglitz (1973), but the following papers are far simpler and more accessible: Stiglitz (1976); King (1975); Flemming (1976). The analysis in the text borrows heavily from Stiglitz (1976) and King (1975).

21. The same result obtains without interest deductibility but with immediate depreciation of full investment costs. In this case, the firm only needs to borrow $(1 - t_c)$ dollars to finance a dollar of additional investment. The remainder can be financed out of tax savings. Hence, the firm's net returns in period t are $(1 - t_c)(\partial\pi/\partial K_t - \delta) - r(1 - t_c)$, with the investment margin again defined by $(\partial\pi/\partial K_t - \delta) = r$, independent of t_c.

$(\partial\pi/\partial K_t - \delta) = r/(1 - t_c)$ on the margin. The cost of capital is directly proportional to increases in t_c, as Harberger intended.

Determining whether or not the tax is actually distortionary would require a full analysis of all its design characteristics, as well as the underlying market environment. For instance, some firms may be subject to borrowing constraints that would change their investment margins. Also, estimated depreciation allowances may not reflect true economic depreciation as assumed above. All things considered, the tax is undoubtedly distortionary to some extent. But in light of Stiglitz's analysis, assuming that the U.S. corporate tax is nondistortionary may be a good approximation to reality.

IMPORTANT MODIFICATIONS OF THE HARBERGER MODEL

The basic Harberger model makes a number of very strong assumptions that need to be modified to extend the usefulness of the model. Three especially useful modifications are variable rather than fixed factor supplies, imperfect rather than perfect competition, and heterogeneous rather than homogenous consumers. We conclude the chapter with brief discussions of each of them.

Variable Factor Supplies

The assumption of fixed factor supplies simplifies an already complex analytical model, but it needs to be modified for the sake of reality. Factor supplies are certainly variable in the long run, with the single possible exception of land, and the supply responses to a tax matter in determining the incidence of the tax. As a general rule, they tend to reduce the change in the price of the taxed factors, which spreads more of the burden to the untaxed factors. For example, a reduction in the supply of capital in the Harberger model would raise the return to capital and thereby transfer some of the tax burden to labor.

The assumption of variable factor supplies is essential in long-run dynamic models of tax incidence, which we will consider in Chapter 17. Changes in the supply of capital in response to a tax alter the time path of capital accumulation, both physical and human capital, which in turn affects the marginal products of capital, labor, and all other factors of production. The resulting changes in marginal products are often the most important determinants of the ultimate incidence of a tax as the economy moves to its new long-run steady state. For example, a tax on capital that slows the rate of capital accumulation and reduces the capital/labor ratio can shift much of the long-run incidence of the tax to labor as the marginal product of labor and therefore the real wage fall over time.

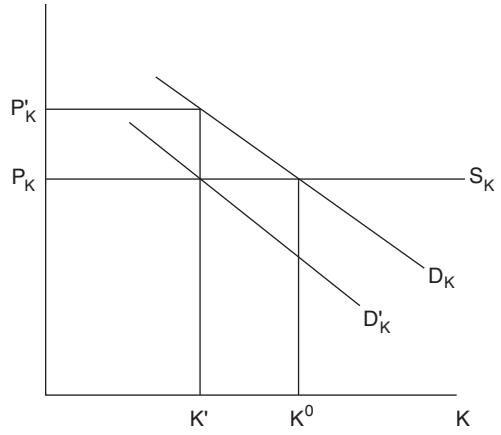

FIGURE 16.8

Variable supply responses to taxation can be important as well in static models of tax incidence. To see this, consider the opposite extreme from the Harberger model. Assume that the supply of capital is perfectly elastic to the taxing jurisdiction, as depicted in Fig. 16.8. This is a realistic assumption for state (provincial) and local governments, and for all but the largest countries with the most highly developed financial markets. The supply price of capital, P_K, is a given for these smaller jurisdictions, determined in capital and financial markets whose scope extends far beyond their boundaries. As a consequence, taxes on capital have very different implications for them than the Harberger model would suggest. The following three implications have received attention in the tax incidence literature.

Mobile versus Immobile Factors

In the first place, a tax on the firms' use of capital in these jurisdictions cannot be borne by capital (refer to Fig. 16.8). The tax shifts D_K down by the full amount of the tax. Since the supply price of capital, the required rate of return, is set on the "world" market, the rate of return within the jurisdiction rises by the full amount of the tax to P'_K. The capitalists escape the burden, passing it on to other factors of production or to consumers. The general principle illustrated here is this: If a jurisdiction contains both mobile factors (e.g., capital) and immobile factors (e.g., land and possible labor), then the immobile factors bear the burden of any factor tax levied on the firms. (If all factors are perfectly mobile, with their prices fixed outside the jurisdiction, then consumers bear the entire burden in the form of higher prices.)

Taxing the Demand versus the Supply Side

Another implication of the perfectly elastic assumption is that it matters which side of the market is taxed. Earlier we

had noted the principle that it did not matter which side of the market the legislature chose to tax: All the effects of a tax on one side of a market can be duplicated by a tax on the other side of the market. The case of perfectly elastic supply is the exception to this principle, and an important one, if the government cannot tax all the suppliers.

Suppose the jurisdiction levies a tax on its own suppliers of capital through, say, a personal income tax. The citizens of the jurisdiction may reduce their saving because the after-tax return to their saving has been reduced. But they represent such a small proportion of the overall supply of saving to the "world" financial markets that they cannot possibly affect the given price of capital, P_K. The tax has no effect whatsoever on the allocation of capital to the jurisdiction. The citizens' after-tax return simply falls by the full amount of the tax, and they bear the entire burden of the tax. The implication is clear. State governments that wish to tax returns to capital in the interests of fairness without harming investment in the state should do so through their personal income tax and not through a corporation income tax or any other business tax on capital. The former has no effect on investment and the incidence sticks where the tax is levied. The latter reduces the equilibrium capital stock, and the burden of the tax is passed on to other factors of production or to consumers.

The Incidence of Local Property Taxes

The supply elasticity of capital has figured prominently in the literature on the incidence of local property taxes in the United States. The local property tax is a combined tax on land and capital. The taxes are levied on the total value of each parcel of land, which includes both the value of the land itself and the value of the structure on the land. The value of the structure is usually much greater than the value of the land, so that the local property tax is primarily a tax on capital.

The original, or "old" view of the incidence of the local property tax had long held that it was a regressive tax overall. The portion of the tax on the value of the land itself is progressive. The supply of land is virtually perfectly inelastic, so that landowners bear the entire burden of this portion of the tax, and the distribution of land ownership is skewed heavily toward high-income households.

The progressivity of the land portion is overwhelmed by the regressivity of the much larger capital portion, according to the old view. The local capital market was seen as in Fig. 16.8, with the supply price of capital given to the locality. The tax on the structures was therefore assumed to be passed on to others by the apartment owners to their renters and by the commercial and industrial firms to their consumers or to labor. The property owners escaped, and the larger portion of the tax was regressive.

The "old" view came to be replaced about 35 years ago by the so-called "new" view, which held that local property taxes were progressive after all. The new view accepts the old view's characterizations of the markets for land and capital. The supply of land is perfectly inelastic, therefore, landowners bear the incidence of the land portion of the tax. The market for capital within localities is as pictured in Fig. 16.8, with the supply of capital perfectly elastic to each locality. What the old view failed to consider, however, is that all localities have property taxes. Therefore, capitalists cannot escape the *average* rate of the property taxes across localities. Since the overall supply of capital in the nation is viewed as fixed, the capitalists bear the average rate of the local property taxes, and the incidence of the capital portion is therefore progressive.

The germ of truth in the old view of the capital market, according to the new view, is that capitalists respond to *differences* in tax rates across localities. Capital would move from jurisdictions with above-average property taxes to jurisdictions with below-average property taxes, until the returns to capital were equalized in all localities. Renters and (immobile) labor in the high-tax jurisdictions lose as the capital leaves, and renters and labor in the low-tax jurisdictions gain as new capital enters. But the differences in local property tax rates are typically much smaller than the overall average tax rates. Also, the incidence resulting from differences in the rates is at least partially a wash, since some renters and consumers gain while others lose. Therefore, the incidence of the capital portion is almost certainly progressive. Since the land portion is also progressive, the local property taxes are progressive, perhaps even highly progressive.

The "new" view of 35 years ago was itself challenged by the advent of dynamic models of tax incidence, which showed that even taxes on land could be regressive. We will return to the property tax in Chapter 17 when we analyze dynamic tax incidence.

Oligopoly and the Corporation Income Tax

Many important industries in the corporate sector in the United States and other developed market economies are better modeled as oligopolies than as perfect competitors (or pure monopolies at the other extreme). This is unfortunate for the study of tax incidence because economists have not been able to develop a general theory of oligopoly, nor are they likely to. Also, the game theoretic approach to oligopoly that has dominated the industrial organization literature for the past 35 years has not paid much attention to taxation.

Nonetheless, the suspicion lingers that the incidence of taxes levied on oligopolies could be quite different from their incidence if levied on perfect competitors or pure monopolies. The primary basis for the suspicion is that

oligopolies are likely to produce more output and charge lower prices than would be required to maximize group profits under an industry cartel. On the one hand, noncooperative strategic considerations in a game theoretic environment may drive them away from the cartel profit maximum. On the other hand, they may simply have different goals in the short run, such as maximizing sales (market share) rather than profits, that lead them to set lower prices.

Suppose the firms are operating away from the cartel profit maximum, for whatever reason. A tax such as a corporation income tax could then cause them to raise prices and restrict output in concert, which moves them closer to the cartel profit maximum. The owners may escape the burden of the tax entirely if they have enough unexploited profits to call upon, such that the after-tax profits with the tax equal the profits without the tax. Indeed, the tax may actually make the firms better off if profits before tax rise by more than the tax liability. This particular escape route is not possible for perfect competitors or profit-maximizing monopolists.

The possibility that oligopolists might escape the burden of taxation in this way was first explored by Harvey Rosen and Michael Katz in 1985 (Katz and Rosen, 1985). They developed a simple conjectural variation model of an industry in which each firm conjectures (guesses) about how the other firms will respond to changes in its prices or output. The conjectures are symmetric—all firms make the same guess, and their model does not allow for entry or exit. One attractive feature of conjectural variation models is that they permit the full range of possible outcomes from the $P = \text{MC}$ perfectly competitive result to the $\text{MR} = \text{MC}$ cartel result, depending on the nature of the conjectures and the number of firms in the industry.

Rosen and Katz illustrated the possibility of escaping the burden of a tax using a very simple example of a duopoly with linear demands and symmetric conjectures. They were particularly interested in the range of consistent, or rational, conjectures, meaning that each firm assumes that the other firm will respond in a profit-maximizing manner to its changes in price or output. They introduced a factor tax that increases marginal cost and found that unexploited profits did rise by more than the tax liabilities under consistent conjectures. The firms more than escaped the tax burden by moving closer to the cartel price and output. Whether actual oligopolies can escape tax burdens in this way remains a wide-open question.

Heterogeneous Consumers

A final important modification of the Harberger model is the assumption of heterogeneous consumers. An obvious implication of heterogeneity is that the incidence of taxation depends on the ownership of factors of production and expenditure patterns across consumers, as well as the amounts by which relative factor and goods prices change.

Although a number of economists have introduced heterogeneity into the Harberger framework, the more recent literature is moving in a different direction. The increase in income inequality in the United States since the mid-1970s (and in many of the other developed market economies) is mostly due to an increase in inequality within earnings and not to an increase in the share of national income going to capital. Therefore, in accounting for heterogeneity, economists have been turning their attention toward the effects of the major taxes on the personal distribution of income. They are no longer so interested in how taxes affect the relative shares of capital and labor income. We will discuss the newer incidence studies in Chapter 17.

REFERENCES

Aaron, H.J., May 1974. A new view of property tax incidence. The American Economic Association Papers and Proceedings 64 (2), 212–221.

Aaron, H.J., 1975. Who Pays the Property Tax? The Brookings Institution, Washington, DC.

Asimakopulos, A., Burbidge, J., June 1974. The short-period incidence of taxation. Economic Journal 84 (334), 267–288.

Auerbach, A., Kotlikoff, L., 1987. Dynamic Fiscal Policy. Cambridge University Press, New York.

Ballentine, J.G., Eris, I., June 1975. On the general equilibrium analysis of tax incidence. Journal of Political Economy 83 (3), 663–644.

Diamond, P., June 1978. Tax incidence in a two-good model. Journal of Public Economics 9 (3), 283–299.

Feldstein, M., October 1974. Incidence of a capital income tax in a growing economy with variable savings rates. Review of Economic Studies 41 (4), 505–513.

Flemming, J., July-August 1976. A reappraisal of the corporation income tax. Journal of Public Economics 6 (1–2), 163–169.

Friedlaender, A.F., Vandendorpe, A., October 1976. Differential incidence in the presence of initial distorting taxes. Journal of Public Economics 6 (3), 205–229.

Fullerton, D., Rogers, D., 1993. Who Bears the Lifetime Tax Burden? Brookings Institution, Washington, DC.

Harberger, A., June 1962. The incidence of the corporation income tax. Journal of Political Economy 70 (3), 215–240.

Kalecki, M., September 1937. A theory of commodity, income and capital taxation. Economic Journal 47 (187), 444–450.

Katz, M., Rosen, H., January 1985. Tax analysis in an oligopoly model. Public Finance Quarterly 13 (1), 3–19.

King, M., August 1975. Taxation, corporate financial policy, and the cost of capital: a comment. Journal of Public Economics 4 (3), 271–279.

Krzyzaniak, M., Musgrave, R.A., 1963. The Shifting of the Corporation Income Tax. Johns Hopkins University Press, Baltimore, MD.

Mieszkowski, P., June 1967. On the theory of tax incidence. Journal of Political Economy 75 (3), 250–262.

Musgrave, R.A., May 1974. Is a property tax on housing regressive? The American Economic Association Papers and Proceedings 62 (2), 222–229.

Pechman, J., 1985. Who Paid the Taxes, 1966–85? The Brookings Institution, Washington, DC.

Pechman, J., Okner, B., 1974. Who Bears the Tax Burden? The Brookings Institution, Washington, DC.

Shoven, J., December 1976. The incidence and efficiency effects of taxes on income from capital. Journal of Political Economy 84 (6), 1261–1283.

Shoven, J., Whalley, J., November 1972. A general equilibrium calculation of the effects of differential taxation of income from capital in the U.S. Journal of Public Economics 1 (3–4), 281–321.

Shoven, J., Whalley, J., October 1977. Equal yield tax alternatives: general equilibrium computational techniques. Journal of Public Economics 8 (2), 211–224.

Stiglitz, J., 1973. Taxation, corporate financial policy, and the cost of capital. Journal of Public Economics 2 (1), 1–34.

Stiglitz, J., April–May 1976. The corporation tax. Journal of Public Economics 5 (3–4), 303–311.

Chapter 17

Expenditure Incidence and Economy-Wide Incidence Studies

Chapter Outline

Taxes are most often raised to finance government expenditure programs, not just to substitute for other taxes. Once this obvious point is conceded, it is no longer as compelling to speak only of the incidence of the tax revenues. The policy-relevant incidence measure is clearly balanced-budget incidence, the entire tax-and-expenditure package. One might still argue that tax incidence itself remains relevant since different sets of taxes could have financed the given expenditure program. Still, ignoring the expenditure side is always dangerous since the very existence of a new expenditure program affects the evaluation of the single tax and differential incidence measures discussed in Chapter 16. Government inputs and outputs enter into the market clearance and

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00017-0

government budget equations, thereby influencing the price responses to any change in tax rates. Also, the distributional consequences of expenditure programs are likely to be as important as the distributional consequences of the tax revenues raised to finance them. Thus, to the extent incidence analysis is an aid to governmental distributional policies, considering the incidence of an entire tax-and-expenditure package would appear to be the most useful strategy.

This is bound to be a difficult assignment, however, even in theory, because balanced-budget incidence theory is fraught with the same difficulties as the theory of tax incidence, plus some other problems as well. At the very least, an analysis of various balanced-budget alternatives must confront these issues at the outset:

1. What measure of incidence will be employed? The three most likely candidates are income compensation or welfare loss measures applied to individuals, the change-in-relative-prices measure in the Harberger tradition, or the change in a many-person social welfare function, the same as for the theory of tax incidence.
2. For any given set of taxes, what expenditure programs are being financed? The obvious candidates are transfer payments, Samuelsonian nonexclusive public goods or other externality-generating goods, and government-operated decreasing-cost services, although the government might be buying goods and services that could have been supplied by a perfectly competitive private sector. Public insurance is another possibility that has not yet been discussed. We can ignore it for now on the grounds that in its pure form it would be paid for on a benefits-received basis, and therefore would not be a candidate for incidence analysis. There are distributional implications in all practical applications, however, which we will consider in Chapters 20 and 21.
3. Will the analysis consider marginal, balanced-budget changes in taxes and expenditures, or must it focus on a total package of finite taxes and expenditures? Marginal analysis might make sense for transfer payments but surely not for decreasing cost services.
4. For any given expenditure program, how are the taxes being raised? The point that the choice of expenditures affects the measurement of tax incidence is reversible. The method of financing the expenditures dictates the approach to the measurement of expenditure incidence. Are the expenditures assumed to be financed with lump-sum taxes or with a set of distorting taxes? If resource-using government expenditure programs are assumed to be financed with lump-sum taxes, then the incidence analysis could take place within a first-best context, so long as other appropriate assumptions are made, such as perfectly competitive private production and marginal-cost pricing of government services. Lump-sum financing would also provide an unambiguous method for considering the incidence of a single (set of) government programs (s), or a separate theory of expenditure incidence, analogous to the incidence of a single-tax program when the revenues are returned lump sum. This is an important consideration, since first-best expenditure incidence is more compelling than first-best tax incidence. Resource-using expenditures are undertaken solely for efficiency reasons in a first-best environment. Nonetheless, they do have distributional consequences, and knowing these aids the government in its search for the optimal pattern of lump-sum redistributions.

If, realistically, governments are assumed to use distorting taxes to finance their expenditures, then the analysis is inherently second best, and tax and expenditure incidence cannot be separated (unless the expenditures happen to be self-financing using benefits-received taxes). As we saw in the discussion of tax incidence in a many-consumer world, one may have no choice but to adopt the aggregate social welfare approach to have a theoretically sound analysis.

We will not attempt an exhaustive analysis of all possible tax-and-expenditure combinations with all possible incidence measures. Rather, we will highlight some of the problems involved with introducing specific expenditure programs into an analysis of incidence. Thus, to keep the discussion manageable, the numerous possibilities will be limited in three ways:

1. Tax-and-expenditure packages will be evaluated by the income compensation or loss measure of incidence.[1] Hence, we will assume a one-consumer-equivalent economy, with an optimal income distribution.
2. Only three expenditure programs will be considered: transfer payments, decreasing costs services, and nonexclusive Samuelsonian public goods.
3. We will assume that lump-sum tax revenues finance the two resource-using expenditure programs—decreasing cost services and Samuelsonian public goods—and analyze their incidence in a first-best environment. Since the theory of resource-using public expenditures in a second-best environment will not be considered until the next chapter, a discussion of second-best expenditure incidence at this point in the text would be premature. In contrast, all the tools necessary for a comprehensive analysis of the incidence of transfer payments in a second-best environment have already been developed.

1. The many-person social welfare measure of incidence will be discussed in Chapter 24. For an analysis of expenditure incidence in the Harberger tradition, see McClure and Thirsk (1975).

THE INCIDENCE OF GOVERNMENT TRANSFER PAYMENTS

Transfer payments, or subsidies, are analytically equivalent to negative taxes. Consequently, the theory of tax incidence is fully applicable to government transfer payments, with the single exception that all signs are reversed. All we need do, then, is review the major results of the previous chapter as they apply to subsidies:

1. If lump-sum taxes finance lump-sum transfers, there is no burden or incidence in a one-consumer-equivalent economy. In a many-person economy, the tax paid or transfer received by any one person will be an appropriate income proxy for the welfare gain or loss by that person under either one of two assumptions: (a) technology is linear so that the taxes and transfers cannot change the equilibrium vector of consumer and producer prices; or (b) the policy-relevant alternative to a given transfer-tax program is for the government to completely undo the program, recalling all transfers and returning all taxes, thereby restoring the original pretax and transfer equilibrium. Otherwise, the tax-transfer program changes relative prices, and an individual's gain or loss would be measured by the value of his expenditure function evaluated at, say, the new prices and original utility level, less the lump-sum tax paid or transfer received.
2. A set of distorting subsidies offered to consumers and financed by lump-sum taxes is formally equivalent to the single-tax incidence problem of levying a set of distorting taxes and returning the revenues lump sum. The distorting subsidy-with-lump-sum tax generates a dead-weight loss measured, in the case of linear technology, by

$$L(\vec{\mathbf{s}}) = M\left(\vec{\mathbf{q}}; \overline{U}^0\right) + \sum_{i=1}^{N} s_i X_i^{\text{comp}} \tag{17.1}$$

where

$\vec{\mathbf{s}}$ = the vector of per-unit subsidies with element s_i,
$\vec{\mathbf{q}}$ = the vector of consumer prices net of subsidy.

The appropriate measure in the case of general technology is

$$L(\vec{\mathbf{s}}) = M\left(\vec{\mathbf{q}}; \overline{U}^0\right) + \sum_{i=1}^{N} s_i X_i^{\text{comp}} - \pi(\vec{\mathbf{p}}) \tag{17.2}$$

The conceptual experiment described by Eqn (17.1) is a comparison between the lump-sum income necessary to reach the original utility level at the new lower prices less the amount of the subsidy, which is returned lump sum, everything measured at the compensated equilibrium. (Equation (17.2) subtracts the pure profits available at the compensated income from the required income.) Hence, Eqn (17.1) measures the payment the consumers are willing to make as a consequence of the subsidies less the required lump-sum income payment (the return of the subsidy).[2]

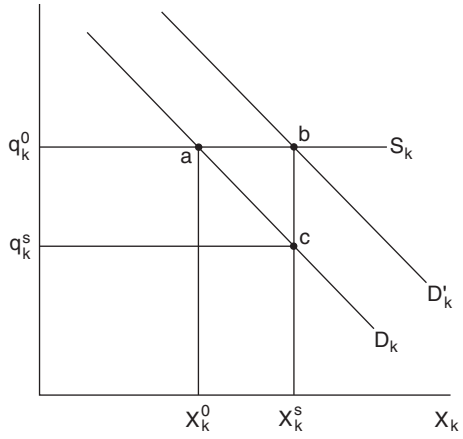

FIGURE 17.1

Figure 17.1 illustrates the measure of loss created by a unit subsidy s_k on the consumption of X_k, with linear technology. The subsidy shifts D_k upward by s_k, reduces the price X_k to the consumer from q_k^0 to q_k^s, and increases consumption from X_k^0 to X_k^s. The gain to the consumer is the area $acq_k^s\, q_k^0$, the area behind the original compensated demand curve (compensated at utility level $\overline{U}^0$) between the old and new prices, Hicks' compensating variation (HCV). The subsidy at the new compensated equilibrium is $s_k\, X_k^s$, the area $bcq_k^s\ q_k^0$, which the consumer must pay for with lump-sum taxes. The net loss, therefore, is the triangle abc. Under the income-compensation measure of incidence, this dead-weight loss is the incidence of the subsidy.[3]

Similarly, the incidence of marginal changes in an entire set of distorting subsidies is measured by summing the changes in dead-weight loss each time, equal (for linear technologies) to

$$\mathrm{d}L = \sum_k \sum_i s_i M_{ik} \mathrm{d}s_k \tag{17.3}$$

where

M_{ik} = the Slutsky substitution terms.

2. Since, from the consumer's point of view, goods prices are falling and factor prices are rising, $M(q; \overline{U}^0)$ measures the income consumers are willing to pay for the subsidies and is a negative number. Hence, loss is the addition of $M(q; \overline{U}^0)$ and $\sum_{i=1}^{N} s_i X_i^{\text{comp}}$, where $s_i > 0$ for goods and <0 for factors. Similarly, goods prices are rising and factor prices are falling from the firm's point of view, both of which tend to increase profits. Hence, $\pi(\vec{\mathbf{p}})$ must be subtracted from the subsidy payment in Eqn (17.2) under general technology. This HCV measure is conceptually equivalent to the HEV measure for a distorting tax.

3. In single-market partial equilibrium analysis, the dead-weight loss from any distortion is always the area between the compensated demand curve and the general equilibrium supply curve measured from the distorted to the undistorted equilibrium quantities.

Likewise, the expression for the total loss from a set of unit subsidies is

$$L = \frac{1}{2}\sum_i \sum_j s_i s_j M_{ij} \tag{17.4}$$

analogous to the total loss from a set of unit taxes.

3. One set of distorting subsidies may be substituted for another while holding the total subsidy constant, a case of differential expenditure analysis. This is exactly analogous to differential tax incidence. Here, the substitution is viewed as removing one set of subsidies, returning the tax savings lump sum, and then instituting a second set of subsidies, paid for by lump-sum taxes. As in the tax case, the first step involves totally differentiating the government's budget constraint:

$$\sum_{i=1}^{N} s_i X_i^{\text{comp}} = \overline{S} \tag{17.5}$$

with $dS = 0$, to determine the changes in the s_i necessary to maintain a balanced budget. The resulting changes are then substituted into Eqn (17.3) to evaluate the change in loss (for linear technologies). Finally, the practical difficulties of applying these compensated measures, especially for general technologies, which were discussed in Chapter 16, apply to transfer incidence as well. Recall that an important issue was whether production is constant returns to scale.

TAX AND EXPENDITURE INCIDENCE WITH DECREASING-COST SERVICES

As long as decreasing-cost services are being analyzed within the context of first-best theory, the government is assumed to charge a price equal to the marginal cost of providing the service and to finance with lump-sum taxes the deficits arising because MC < AC. The appropriate comparison is an all-or-none test in which having the service with these characteristics is compared to not having the service at all. Marginal incidence analysis is not relevant for decreasing-cost services.

The income-compensation measure of incidence was developed in Chapter 9. Assuming linear or constant-returns-to-scale (CRS) general production technology elsewhere in the economy, the net benefit of providing the decreasing-cost service with lump-sum financing of its deficit is[4]

$$B = -M\left(\vec{\mathbf{q}}; \overline{U}^0\right) - T \tag{17.6}$$

where

$\vec{\mathbf{q}}$ = vector of consumer prices with the service,
$\overline{U}^0$ = the utility level without the service,
T = the lump-sum payment required to finance the deficit,
$-M(\vec{\mathbf{q}}; \overline{U}^0)$ = the amount consumers are willing to pay for the new prices, $\vec{\mathbf{q}}$.

SAMUELSONIAN NONEXCLUSIVE GOODS

Chapter 6 developed the standard pareto-optimal decision rule for a nonexclusive good in a first-best environment, $\sum_{h=1}^{H} \text{MRS}^h = \text{MRT}$, but did not consider the incidence of the good. The incidence is the gain in welfare to each consumer from being able to consume the good at its optimal level, less the loss in welfare from having to finance the good.

As a first step in deriving an incidence measure, recall that all government decisions with respect to financing and providing the good are lump-sum events from any one consumer's point of view. Since the market system completely breaks down because of the revelation problem, the government has no choice but to select a given quantity of the good that will be available in equal amounts to all consumers, hope that it satisfies the $\Sigma\text{MRS} = \text{MRT}$ rule, and then finances its purchases with lump-sum taxes to preserve efficiency in all other markets.

For the purposes of this discussion, assume that the government has selected the optimal quantity, so that $\sum_{h=1}^{H} \text{MRS}^h = \text{MRT}$. Assume further that production of the nonexclusive good and all other goods and services exhibits either CRS or linear technology.

Consumers react in two ways to the existence of a nonexclusive good. On the one hand, the good enters each consumer's utility function directly as one of the arguments, although the sign of the argument is uncertain. Some consumers may view it as a "good," others as a "bad," especially at the margin. On the other hand, consumers may well adjust their own goods demands and factor supplies in response to the nonexclusive good. That is, the nonexclusive good may be a substitute for or complement to other goods and factors.

A representation of the consumer's indirect utility function that captures these features is

$$V\left(\vec{\mathbf{q}}; \overline{I}; e\right) = U\left[X_i\left(\vec{\mathbf{q}}; \overline{I}; e\right); e\right] \tag{17.7}$$

with

$$\frac{\partial V}{\partial e} = \sum_{i=1}^{N} \frac{\partial U}{\partial X_i}\frac{\partial X_i}{\partial e} + \frac{\partial U}{\partial e} \tag{17.8}$$

4. Alternatively,

$$B = \left[M\left(\vec{\mathbf{q}}^0; \overline{U}\right) - M\left(\vec{\mathbf{q}}; \overline{U}\right)\right] - T, \quad \text{with } M\left(\vec{\mathbf{q}}; \overline{U}\right) = 0 \tag{17.6N}$$

The term in brackets is HCV measure of the willingness to pay for the price change, where $\vec{\mathbf{q}}^0$ = the vector of consumer prices without the service. Alternatively, Hicks' equivalent compensation could be used, comparing the before and after prices at the with-service utility level.

where

$\vec{\mathbf{q}}$ = the vector of consumer prices,

X_i = good (factor) i demanded (supplied) by the consumer,

$\overline{I}$ = a source of lump-sum income other than profits from production, assumed constant unless taxed by the government,

e = the quantity of the nonexclusive good selected by the government.

Two results useful for the measure of incidence follow directly from the first-order conditions of utility maximization. First, differentiate the budget constraint with respect to e to obtain

$$\sum_{i=1}^{N} q_i \frac{\partial X_i}{\partial e} = 0 \tag{17.9}$$

From the primal of the consumer problem,

$$\frac{\partial U}{\partial X_i} = \lambda q_i \quad i = 1, \ldots, N \tag{17.10}$$

Substituting Eqn (17.10) into (17.9) yields

$$\frac{1}{\lambda} \sum_{i=1}^{N} U_i \frac{\partial X_i}{\partial e} = 0 \tag{17.11}$$

Thus, Eqn (17.8) simplifies to

$$\frac{\partial V}{\partial e} = \frac{\partial U}{\partial e} \tag{17.12}$$

The change in utility from a marginal change in the nonexclusive good equals its direct marginal effect on utility. Although consumers may change their other purchases and factor supplies in response to the change in e, these changes have no further effect on utility.

Second, Eqn (17.12) implies that the marginal rate of substitution between e and ith good or factor, MRS_e, X_i, is defined exactly as it would be for any exclusive good:

$$\mathrm{MRS}_e, x_j = -\frac{\frac{\partial U}{\partial e}}{\frac{\partial U}{\partial X_i}} \tag{17.13}$$

If good i is the numeraire, then

$$\mathrm{MRS}_e, x_j = -\frac{1}{\lambda}\frac{\partial U}{\partial e} - \frac{\mathrm{d}U}{\mathrm{d}e} \Big/ \frac{\mathrm{d}U}{\mathrm{d}I} = -\mathrm{d}I/\mathrm{d}e_{U=\overline{U}} \tag{17.14}$$

Thus, the marginal rate of substitution establishes the value of a marginal increase in the public good to the consumer, as it does for any good.

The value of a finite amount of the public good can be derived from the consumer's expenditure function. In the presence of a nonexclusive good, the dual to the standard consumer problem is

$$\min_{(X_i)} \sum_{i=1}^{N} q_i X_i$$
$$\text{s.t. } U = \overline{U}\left(\vec{\mathbf{X}}; e\right)$$

The first-order conditions yield compensated demand (supply) functions of the form

$$X_i^{\mathrm{comp}} = X_i\left[\vec{\mathbf{q}}; \overline{U}\left(\vec{\mathbf{X}}; e\right)\right] \quad i = 1, \ldots, N \tag{17.15}$$

and the expenditure function:

$$M\left[\vec{\mathbf{q}}; \overline{U}\left(\vec{\mathbf{X}}; e\right)\right] = \sum_{i=1}^{N} q_i X_i^{\mathrm{comp}}\left[\vec{\mathbf{q}}; \overline{U}\left(\vec{\mathbf{X}}; e\right)\right] \tag{17.16}$$

Thus, even though the consumer does not purchase e, the expenditure function has e as an argument because e appears in the utility function, which is being held constant. All we need establish, then, is that $\partial M/\partial e \neq 0$, so that as e changes the income required to keep the consumer at the same utility level also changes:

$$\frac{\partial M}{\partial e} = \sum_{i=1}^{N} q_i \frac{\partial X_i^{\mathrm{comp}}\left[\vec{\mathbf{q}}; \overline{U}\left(\vec{\mathbf{X}}; e\right)\right]}{\partial e} \tag{17.17}$$

Substituting Eqn (17.10) into (17.17) yields

$$\frac{\partial M}{\partial e} = \frac{1}{\lambda} \sum_{i=1}^{N} \frac{\partial U_i}{\partial X_i} \frac{\partial X_i^{\mathrm{comp}}\left[\vec{\mathbf{q}}; \overline{U}\left(\vec{\mathbf{X}}; e\right)\right]}{\partial e} \tag{17.18}$$

But $U = \overline{U}(\vec{\mathbf{X}}; e)$. Thus,

$$\sum_{i=1}^{N} \frac{\partial U_i}{\partial X_i} \frac{\partial X_i^{\mathrm{comp}}}{\partial e} + \frac{\partial U}{\partial e} = 0 \tag{17.19}$$

if utility is held constant, or

$$\sum_{i=1}^{N} \frac{\partial U_i}{\partial X_i} \frac{\partial X_i^{\mathrm{comp}}\left[\vec{\mathbf{q}}; \overline{U}\left(\vec{\mathbf{X}}; e\right)\right]}{\partial e} - \frac{\partial U}{\partial e} \tag{17.20}$$

Hence,

$$\frac{\partial M}{\partial e} = -\frac{1}{\lambda}\frac{\partial U}{\partial e} = -\left.\frac{\mathrm{d}I}{\mathrm{d}e}\right|_{U=\overline{U}} \tag{17.21}$$

As expected, the derivative of the expenditure function with respect to the nonexclusive good yields the change in lump-sum income that makes the consumer indifferent to a change in the nonexclusive good. From Eqn (17.19), this is nonzero, in general. Also, from Eqn (17.14), $\partial M/\partial e$ is the marginal rate of substitution between e and the numeraire good.

An appropriate income measure of the gain from having a finite amount of a nonexclusive good is[5]

5. If I varies as e varies because of general technology, then,

$$B = \left[(I_e - I_0) + M(\vec{\mathbf{q}}^0; \overline{U}^0)(e = 0) - M\left(\vec{\mathbf{q}}^0; \overline{U}^0(e)\right)\right] \tag{17.22N}$$

The gain equals $(I_e - I_0)$, the actual change in lump-sum income as e moves from 0 to e, plus the amount the consumer is willing to pay to have e increased from 0 to e. Recall that $I_0 = M\ M[\vec{\mathbf{q}}^0; \overline{U}^0]$. Therefore, $B = I_e - M[\vec{\mathbf{q}}^0; \overline{U}^0(\vec{\mathbf{X}}; e)]$, Eqn (17.22).

$$B = M\left[\vec{\mathbf{q}}; \overline{U}^0(\vec{\mathbf{X}}; e = 0)\right] - M\left[\vec{\mathbf{q}}; \overline{U}^0(\vec{\mathbf{X}}; e)\right] = \overline{I} - M\left[\vec{\mathbf{q}}; \overline{U}^0(\vec{\mathbf{X}}; e)\right] \tag{17.22}$$

where

$\vec{\mathbf{q}}$ = the vector of consumer prices in the presence of the nonexclusive good,
$\overline{U}^0$ = the consumer's utility when $e = 0$,
$\overline{I}$ = the consumer's lump-sum income, assumed constant.

Notice that Eqn (17.22) would measure the benefit (harm) of any lump-sum event that affects the consumer.

If consumers are asked to make a lump-sum tax payment to finance e or changes in e, then the incidence of the entire tax−income−expenditure package is straightforward. Since the expenditure function expresses welfare changes in terms of lump-sum income, the lump-sum tax is just subtracted from Eqn (17.22) to obtain the incidence of the entire package. Thus, the net benefit is

$$B^N = -M\left[\vec{\mathbf{q}}; \overline{U}^0(\vec{\mathbf{X}}; e)\right] + (\overline{I} - T) \tag{17.23}$$

For marginal charges,

$$\frac{\partial B^N}{\partial e} = -\frac{\partial M}{\partial e} - \frac{\partial T}{\partial e} = \text{MRS}_{e,\text{numeraire}} - \frac{\partial T}{\partial e} \tag{17.24}$$

where

$\frac{\partial T}{\partial e}$ = the change in lump-sum taxes per unit change in e.

Equation (17.24) establishes the following result. Suppose the government is able to establish Lindahl prices for each person equal to the MRS, in accordance with the competitive interpretation of the benefits-received principle of taxation as discussed in Chapter 6. If $\Sigma\text{MRS} = \text{MRT}$ and all production exhibits CRS, the Lindahl prices are sufficient to cover the full costs of the public good. Lindahl pricing also guarantees positive net benefits to all consumers as long as the MRS declines as e increases or the compensated demand for e is downward sloping. Even consumers who think e is a "bad" on the margin gain net benefits with Lindahl pricing. Since their $\text{MRS} > 0$, they would receive subsidies, and these per-unit subsidies would be greater than required on the inframarginal units of e (their MRS is increasing in e). On the margin, however, changes in e accompanied by Lindahl prices generate no net benefits or losses. This is true for any good for which the price equals its MRS (defined in terms of the numeraire good).

The Incidence of Nonexclusive Goods: Empirical Evidence

Although theoretical formulas for the total or marginal incidence of nonexclusive goods are easy enough to derive, they are always very difficult to apply in practice. The problem is the familiar one that consumers have no incentive to reveal their true demand for nonexclusive goods. In particular, we saw that the marginal benefit to the consumer, $\frac{dM}{de} = -\frac{dI}{de}\Big|_{U=\overline{U}}$, is the marginal rate of substitution between the nonexclusive good and an exclusive numeraire good. Yet, no market or political mechanism exists through which the government can accurately measure each consumer's MRS, and incentive-revealing schemes such as Clarke taxes have never been used. Thus, empirical analysis must resort to indirect methods to determine the incidence of these goods.

Researchers often use extremely simple rules to allocate the benefits of public goods such as defense for want of any better alternatives. Examples include allocating the total expenditures per person or per family or in proportion to income per person or per family. In 1970, Henry Aaron and Martin McGuire published an attempt to go beyond these simple allocation methods by incorporating the $\Sigma\text{MRS} = \text{MRT}$ pareto-optimal rules. Even so, they were forced to make some extremely strong assumptions, most notably that everyone has the same preferences and that utility is additively separable in defense. Therefore, everyone is assumed to derive the same utility from defense expenditures. Also, they based the incidence of the public good, e, on the "pseudo-market value" of the good, $\text{MRS} \cdot e$, rather than making use of the expenditure function.[6]

A modest literature proposing different methods for distributing the benefits of public (nonexclusive) goods evolved in response to Aaron and McGuire. The proposals are motivated by the following problem with Lindahl pricing. Let p be the price of a private composite commodity, y_i be the after-tax income of consumer i, and p_i^e be the Lindahl price paid by consumer i. Ask what amount of lump-sum income, M_i, person i would require under Lindahl pricing to be indifferent to receiving the public good free of charge. The required M_i is the solution to $V(p, p_i^e, y_i + M_i) = U(y_i, e)$, where $V(\ )$ is the indirect utility function, $U(\ )$ is the direct utility function, and y_i is spent on the composite commodity. M_i is the benefit received by i for public good e.[7]

The problem is that p_i^e are endogenous; they depend on consumers' tastes and incomes. Consequently, with heterogeneous consumers facing different p_i^e, the M_i are generally not comparable across consumers as income

6. The seminal article is Gillespie, 1965. Gillespie uses a number of simple allocation formulas for different kinds of public expenditures. Aaron and McGuire, 1970. See also the public choice view represented by Maital, 1975. Notice that the MRS for e will vary across people with different incomes even though the total utility they receive from e is the same in the Aaron−McGuire model.

7. In the Lindahl equilibrium with constant returns to scale, the M_i are the pseudo-market values of the good, $p_i^e e$.

measures of utility differences. Thus, the goal became to develop some method of standardizing the pseudomarketing of the public good so that the M_i are comparable utility compensation measures.

James Hines has one of the latest proposals, in the form of a linear pricing scheme. Suppose each consumer could purchase e at the same (linear) price p_e. Ask, as above, what M_i would set $V(p, p_e, y_i + M_i) = U(y_i, e)$, with p_e set such that $\sum_{i=1}^{H} M_i = e$. That is, the sum of the benefits equals the cost of supplying e. The M_i less the actual taxes paid by consumer i to finance e equal the net fiscal benefit (burden) of the public good. The benefits M_i under this proposal are valid income measures of the difference between the utility each consumer would receive if required to purchase e at p_e and the utility at the e chosen by the government (and offered free of charge). Put differently, the benefits defined by the M_i would allow each consumer to achieve the utility at the actual e if instead they were required to purchase e at a common price. Hines argues that taxes set according to M_i would define taxation according to the benefits-received principle in a manner most closely imitative of the usual single-price market mechanism. It is a cost-based mechanism for distributing the benefits, not a surplus-based mechanism, just as is the pseudo-market value at Lindahl prices.[8]

Applying his method to a sample of US households, Hines finds that the M_i first rises and then falls with income when the direct utility function is assumed to be Cobb–Douglas. The benefits are low for low-income consumers because they place a relatively low value on e. The benefits are low again for high-income consumers because their desired e at p_e is far removed from the actual e provided. Since actual federal income tax payments are progressive, the net fiscal burdens are highly progressive at the higher income levels. Aaron and McGuire's pseudo-market value measure of benefits produced a much less progressive pattern of net fiscal burdens.

Whether any such counterfactual cost-based experiments for distributing public goods benefits are persuasive is undoubtedly a matter of taste. In any event, no consensus has been achieved in measuring the incidence of public goods. As we shall see in the remainder of the chapter, economy-wide studies of incidence that include a public good often focus exclusively on the tax side of the government budget. Two approaches to the incidence of the public good are commonplace. One is to simply ignore the effect of the public good on utility. The other is to adopt the Aaron–McGuire assumption of identical preferences with additively separable public goods, and then argue either that the commonly provided public good cannot lead to a difference in welfare across individuals, or that a unit of public good is equivalent to a unit of lump-sum income to each consumer. Neither position is consistent with the expenditure function measure of individual welfare, $M[\vec{\mathbf{q}}; \overline{U}(\vec{\mathbf{X}}; e)]$ without further simplifying assumptions on M.

ECONOMY-WIDE INCIDENCE STUDIES

In 1980, Alan Blinder published a study of the personal distribution of income in the United States covering the 30-year period from 1947 to 1977 (Blinder, 1980); 1947 was the year that the federal government began collecting data on the personal distribution. Blinder's main conclusion was that the distribution was essentially unchanged during those 30 years, a conclusion that surprised him given the economic and demographic turmoil during those years and the rapid growth of the government sector into domestic areas.

The timing of Blinder's study was somewhat ironic, because subsequent research revealed that the personal distribution of income started to become more unequal sometime in the mid-to-late 1970s, a trend that continued at least until 1994, when it appeared to have stopped, only to resume again in the 2000s. Roughly speaking, the families and individuals at the top of the distribution gained at the expense of those at or near the bottom of the distribution in the 1970s and 1980s. From the early 1990s on, those at the top have gained relative to everyone else.

The two very different trends in the personal distribution of income in the last half of the century ignited a huge body of research on the determinants of the distribution. Public sector economists have contributed to this research agenda with a variety of economy-wide tax incidence studies that attempt to measure the impact of the five major US taxes on the personal distribution.[9] They are, in descending order of importance (dollars in billions, 2011, the last year data are available for state and local governments when this was written): federal and state personal income taxes ($1376), federal payroll tax that finances the Social Security system ($819), general sales (state and local) and selective excise (all governments) taxes ($533), local property taxes ($443), and federal and state corporation income taxes ($229).[10]

8. Hines (2000). Hines also demonstrates that, under his proposed linear pricing scheme, the consumers' incentives are to have the government set p_e such that the allocation of e is optimal.

9. The incidence of public expenditures has not received the same attention, with the exception of Social Security and public assistance transfer payments. Similar trends in the distribution in the other developed market economies have led to the same kinds of incidence studies of their major taxes.

10. A few cities also levy personal income taxes, and the revenues from these taxes are included in the data. The sales and excise tax revenues consist of $301 billion from general sales taxes ($236 billion state, $65 billion local) and $232 billion from selective excise taxes ($131 billion states, $72 billion federal, and $28 billion local). General sales and excise taxes tend to be treated similarly in economy-wide incidence studies. The property tax data include revenues from the few states that levy property taxes. The state and local tax data are from Barnett and Vidal (2013). The federal tax data are from the *Budget of the United States Government, Fiscal Year 2014*.

The tax incidence studies have for the most part employed one of three quite different modeling strategies: a heuristic sources and uses approach, computable general equilibrium (CGE) models, and dynamic models of tax incidence. Each strategy has its strengths and weaknesses. None is entirely convincing as a model for determining the overall incidence of a nation's tax system.

We will begin with the sources and uses approach because it appeared first in the literature. Indeed, it was the only method available to economists for economy-wide incidence analysis until the 1970s, when advances in computer technology and computing algorithms gave birth to the other two methods.

THE SOURCES AND USES APPROACH

The sources and uses approach to economy-wide tax incidence is essentially ad hoc. The incidence of the five major taxes is determined by a set of assumptions that allocate the burdens to either the sources or the uses of income. The assumptions pay some attention to general equilibrium tax theory, but only some. The approach accepts the principles that individuals bear the burden of taxation and that the changes in prices in response to taxation ultimately determine the incidence of the tax. At the same time, however, it ignores the dead-weight loss arising from taxation. The approach also considers the incidence of each tax in isolation, ignoring potential interdependencies among the taxes. Finally, the alternative against which the current tax system is being compared is usually treated casually. The general equilibrium with the current tax system should be compared with the general equilibrium that would exist with an alternative tax system that raises the same amount of revenue. The usual comparison, however, is with an equal-revenue, single-rate comprehensive income tax, which is simply assumed to generate a pattern of tax burdens proportional to income. This may not be true. For instance, a flat-rate income tax could generate very unequal dead-weight losses per dollar of revenue in the markets for labor and capital if the (compensated) supply and demand elasticities for labor and capital are quite different.

Annual Incidence Studies

The first sources and uses studies adopted an annual perspective on tax incidence. Lifetime incidence studies did not appear until the 1980s. The sources of personal income on an annual basis are transfer receipts (mostly public), income from labor, and income from capital (income from land is inconsequential). The uses of income are consumption and saving. The goal in allocating the tax burdens to these sources and uses is a modest one, to give a rough idea of whether the overall tax system is progressive, proportional, or regressive in terms of individuals' or families' comprehensive income. This is the best the method can hope to achieve given the ad hoc nature of its assumptions.[11]

The "play" in the assumptions about the allocations of annual tax burdens comes from transfers and income from capital on the sources side and from the uses of income. Government transfer receipts are concentrated among the elderly and the low-income young. Therefore, allocations of tax burden to transfer income tend to be regressive. Conversely, income from capital is highly concentrated among the wealthy. The Gini coefficient for holdings of financial wealth in the United States is on the order of 0.80, compared with an annual income Gini coefficient of approximately 0.48 (2012). Therefore, allocations of tax burden to income from capital are highly progressive. On the uses side, the ratio of annual consumption to income falls sharply as income rises; therefore, allocations of tax burden to consumption tend to be highly regressive.

The distribution of income from labor tends to follow the overall personal distribution of income. Thus, the incidence of tax burdens allocated to labor income depends largely on the structure of the particular tax, for example, the nature of its exemptions, exclusions, or deductions from the tax base and whether it has graduated tax rates.

The Pechman–Okner Studies

The two most widely cited annual sources and uses incidence studies were produced by the Brookings Institution. The first was by Joseph Pechman and Benjamin Okner, published in 1974. The second was an update of the first study in 1985, authored by Pechman alone (Pechman and Okner, 1974; Pechman, 1985). Pechman and Okner merged the panel data on families and individuals from the Survey of Income and Education with Internal Revenue

11. The evolution of the sources and uses approach followed the development of the large micropanel data sets that began to appear in the 1970s. Public sector economists had two choices prior to the micro-data sets. They could construct "typical" families with given incomes, expenditure patterns, and sizes and use the tax laws and price-shifting assumptions to allocate the burdens of the five taxes to these constructed families. Alternatively, they could allocate the aggregate tax revenues on the basis of the price-shifting assumptions to income classes broadly defined, such as by deciles or quintiles. The micro-data sets permitted the allocation of the tax burdens to tens of thousands of actual families and unrelated individuals on the basis of their personal incomes and other characteristics. Some studies went even further, merging IRS data on tax revenues collected from individuals and families with data on individuals and families from one of the panel data sets to get an even more accurate picture of microlevel tax burdens. For further discussion on the evolution of the data used in tax incidence studies, see Atkinson, 1994. Atkinson also has further discussion and analysis of the appropriate alternative against which to compare the current tax system.

Service (IRS) tax files on individual taxpayers to compile a massive data set on incomes, expenditures, personal characteristics, and tax liabilities for US families and unrelated individuals.

Central-Variant Assumptions

Pechman and Okner's most preferred or "central-variant" assumptions for allocating the burdens of the five taxes are as follows.

Personal Income Taxes

They assume that the impact equals the incidence, on the grounds that approximately 80% of the tax base is labor income and that the overall supply of labor was assumed at the time to be almost perfectly inelastic. Their assumption that the tax liability is a reasonable measure of the burden ignores the differences in supply elasticities among men and women that empirical analysis has uncovered over the past 30 years. It also ignores the substantial amount of dead-weight loss from the federal and state income taxes that was discovered by Hausman and others. Recall that the loss is due to a substantial substitution effect that is offset by an income effect of approximately equal magnitude (thereby accounting for the near-zero actual supply elasticity).

Allocating the tax burden by tax liability leads to the conclusion that the personal income taxes are fairly progressive overall and highly progressive at low incomes. The overall progressivity is due to the graduated federal tax rates, and the steep low-end progressivity is due to the personal exemptions and the standard deduction in the federal tax. (Many state income taxes incorporate these same features.)

Payroll Tax for Social Security

Labor is assumed to bear the entire burden of the payroll tax, even though Congress levies half the tax rate on the employers and half on the employees. Also, the tax burden equals the total tax liability, as with the personal income tax. These assumptions follow because the payroll tax is levied only on wage and salary income, and the overall supply of labor is assumed to be perfectly inelastic.[12]

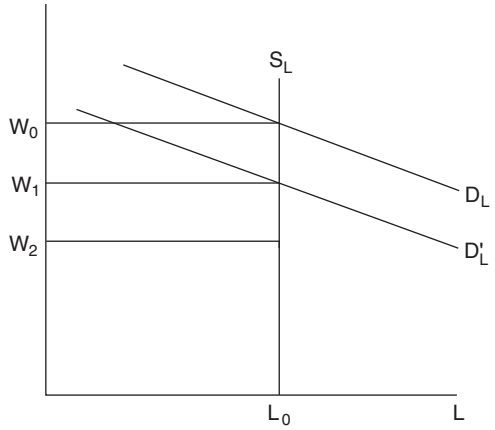

FIGURE 17.2

Figure 17.2 illustrates this point. The equilibrium without the tax is (L_0, W_0). The half of the tax levied on the employers shifts the demand for labor D_L down by the full amount of their tax liability to D'_L. With S_L perfectly inelastic, the wage falls to W_1. The employers fully escape the tax by lowering the wage against the perfectly inelastic supply. Then, the half of the tax on the employees reduces their after-tax wage to W_2, again by the full amount of their liability. Therefore, the employees bear the entire burden of the payroll tax, equal to the combined tax liabilities of the employers and the employees.

The payroll tax is highly regressive under these assumptions because the tax was levied at a flat rate on wage and salary income below a maximum income limit in 1974 and 1985. Income above the maximum was untaxed. Therefore, a person earning $200,000 or $2,000,000 per year paid the same tax as another person earning the maximum amount of taxable income.[13]

Sales and Excise Taxes

Pechman and Okner assume that the markets for goods and services are perfectly competitive, and that the long-run supply curves are perfectly elastic at constant average and marginal costs. Therefore, prices rise by the full amount of the taxes in the long run, so that the burdens are allocated on the uses side on the basis of consumption. The sales and excise taxes are highly regressive under these assumptions on an annual basis. They tend to be flat-rate taxes such that tax collections are roughly

12. The Social Security system operated strictly on a pay-as-you-go basis until 1983, in which all taxes collected from current employers and employees were paid out to current retirees. There was no explicit benefits-received link between the tax payments and future pension benefits, although the promise of future benefits was certainly implicit. Reforms in 1983 allowed for an accumulation of some of the tax revenues in a trust fund that would eventually finance the retirement benefits of the Baby Boom generation. Even so, viewing the payroll tax as a benefits-received tax would require a lifetime perspective, not an annual perspective.

13. In 1994, Congress removed the income limit from the Medicare portion of the payroll tax. The Medicare portion now applies to all wage and salary income earned in occupations covered by Social Security. The tax is still highly regressive under the Pechman–Okner assumptions, however, as the Medicare portion is only 2.9 percentage points of the combined 15.3% tax rate (in 2014). The Medicare portion will rise over time, however, making the payroll tax less regressive. The maximum income against which the remaining 12.4% applies was $117,000 in 2014.

TABLE 17.1 Effective Average Tax Rates by Income Deciles Under the Central-Variant Assumptions

	Decile									
Effective tax rate (%)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	20.6	20.7	21.1	22.3	23.4	23.8	24.2	25.5	26.4	27.1

Source: Adapted from J. Pechman, *Who Paid the Taxes, 1966–85?*, The Brookings Institution, Washington, DC, 1985, Table 4.4, p. 48.

proportional to consumption, and the ratio of annual consumption to income falls sharply as income rises.[14] The 10% of the population with the highest incomes account for about 80% of the total personal saving in the United States.

Local Property Taxes

Pechman and Okner adopt the new view of the property tax, which assumes that the vast proportion of the property tax burden falls on the owners of land and capital. They assign the entire burden to land and capital in their central variant, which makes these taxes highly progressive.

Corporation Income Taxes

Pechman and Okner accept Harberger's view that the corporation income tax is borne by capital as their most preferred assumption and assign the entire burden to income from capital in their central variant. These taxes are therefore highly progressive. Notice that the Harberger model, which has only one kind of capital, implicitly assumes that all forms of capital are perfect substitutes in production and thus earn the same after-tax rate of return.

The Pechman–Okner central-variant assumptions led them to essentially the same conclusions about the incidence of the overall US tax system in 1974 and 1985. Table 17.1 presents their central-variant tax burdens as a proportion of income by deciles in 1980, reported in Pechman's 1985 study. The overall incidence is mildly progressive throughout the distribution. The mildly progressive personal income taxes and the highly progressive property and corporation taxes slightly dominate the highly regressive payroll and sales and excise taxes. The Pechman–Okner central variant has become the consensus view of the incidence of the US tax system. For instance, it is the one reported in most of the Principles textbooks written for the US market.

14. Twenty-six of the 44 states that levy general sales taxes exempt food purchased for home consumption from the tax bases to lessen the perceived regressivity of these taxes. Even so, the state sales taxes are highly regressive under the Pechman–Okner assumptions.

Alternative Assumptions

Pechman and Okner present a number of alternative assumptions concerning the local property and corporation income taxes that might be plausible from an annual perspective.

Local Property Taxes

The new view of property tax incidence argues that some of the incidence of the tax could be passed on to nonmobile labor or renters as capital moves in response to differences in the effective tax rates across localities. In light of this argument, Pechman and Okner provide an alternative allocation in which 1/2 of the property tax burden is allocated to capital income, 1/4 of the burden is allocated to labor income, and 1/4 of the burden is allocated to consumption of housing services. This allocation makes the property tax much less progressive than the central variant, in which all of the burden is borne by land and capital. It also seems excessive, however, as nonmobile workers and renters gain in the localities with lower than average property tax rates.

Another possibility is to argue that the local property tax is simply a benefits-received tax, a price that the citizens pay for the locally provided public services. Benefits-received taxes are never part of an incidence calculation because they cannot be a net burden. This assumption is also extreme. It might apply in a frontier environment in which people are highly mobile and towns are continually forming and reforming, expanding and contracting. In such a world, people of like tastes for public services would join together, form a town, provide exactly the public services they want, and levy taxes to pay for the public services.[15] The taxes would be benefits-received taxes. But, in a more realistic setting of fixed communities and limited mobility, most people are unlikely to obtain their most preferred bundle of public services. If not, then their property taxes are not benefits-received taxes, and the incidence of the property taxes remains a relevant question. In any event, viewing the local property

15. Chapter 27 presents a formal model of jurisdiction formation in a frontier environment.

tax as a benefits-received tax would also make the overall US tax system less progressive.

Corporation Income Taxes

The implicit Harberger assumption that all forms of capital are perfect substitutes in production and therefore earn the same rate of return is fairly extreme. An alternative is to concede that there is some segmentation among capital markets. One obvious possibility that Pechman and Okner consider is a segmentation between housing and other kinds of physical capital: They may not earn the same after-tax returns because they serve such different purposes. This distinction, if applicable, is particularly relevant for the United States because housing is taxed much more lightly than other forms of capital. Pechman and Okner allocate corporation income taxes to dividends rather than all income from capital under this assumption, which makes the corporation income taxes slightly more progressive than under the central variant.

Another possibility is to assume that corporations are profit satisficers rather than profit maximizers and have leeway to pass the tax on to consumers in the forms of higher prices and reduced output. This is the avenue of escape modeled by Rosen and Katz and discussed in Chapter 16. The ability to pass the tax forward to consumers changes the corporation income taxes from highly progressive taxes under the central variant to highly regressive taxes on an annual basis.

Yet another possibility is to assume that the corporation income tax is essentially a benefits-received tax, a special levy on corporations that the stockholders pay for the privilege of limited liability against losses. The corporation income tax drops out of the incidence calculations under this view, and the overall tax incidence becomes less progressive.

The various alternative assumptions concerning the local property and corporation income taxes can change the incidence of these taxes rather dramatically. But they still do not have much effect on the Pechman–Okner central-variant pattern of tax incidence in the United States because they are the smallest of the five main taxes, much less important than the personal income, payroll, and sales and excise taxes. Researches would have to make very different assumptions about the incidence of the other three taxes to have a substantial impact on the central-variant pattern. Quite different assumptions are possible for the other taxes, but they require a change from an annual to a lifetime perspective. Many public sector economists, perhaps even the majority, would favor switching to a lifetime perspective on tax incidence, and the sources and uses literature has been moving in the direction of lifetime incidence over the past 15–20 years.

Mixing Annual and Lifetime Incidence

The first break from the annual perspective came from Edgar Browning and William Johnson in 1979. They used a lifetime perspective to argue that the incidence of sales and excise taxes is likely to be slightly progressive rather than highly regressive, as was universally assumed at the time (Browning and Johnson, 1979).

Their argument begins with two facts about the US economy. First, the present value of lifetime consumption is approximately proportional to the present value of income for the vast majority of people. Only a small percentage of people leave substantial bequests or, conversely, receive substantial inheritances. Second, many government transfer payments are indexed to the consumer price index (CPI), including Social Security benefits and, implicitly, in-kind public assistance such as food stamps and medical care. These two facts considerably alter the implications of assuming that these taxes are passed on to consumers through higher prices. On the one hand, allocating the tax burden in proportion to consumption is equivalent to allocating the tax burden in proportion to income from a lifetime perspective. On the other hand, transfer income is protected from the tax burden by the CPI indexing. Therefore, the tax burden should be allocated in proportion only to earned income from a lifetime perspective. This implies that the incidence of sales and excise taxes is slightly progressive because transfer income is received disproportionately by the poor. Browning and Johnson conclude that the incidence of the US tax system is somewhat more progressive than suggested by Pechman and Okner's central-variant assumptions.

Whalley's Critique of the Sources and Uses Approach

John Whalley offered a blistering critique of the sources and uses approach to tax incidence in his 1984 Presidential address to the Canadian Economic Association (Whalley, 1984). The fundamental weakness in the approach, according to Whalley, is its reliance on ad hoc assumptions and theorizing. By suitably mixing assumptions from annual and lifetime perspectives and selectively borrowing from tax incidence theory, researchers can generate almost any result they want.

In particular, Whalley applied the Pechman–Okner central-variant assumptions to the Canadian tax system, which is similar to the US tax system, and produced essentially the same mildly progressive annual pattern of incidence that Pechman and Okner found for the United States. The burden of the Canadian taxes by income deciles ranged from 27.5% at the low end to 43% at the high end, with most of the other deciles in the 30–40% range. By selectively varying the assumptions, however, Whalley was

able to generate either a steeply progressive overall tax burden ranging from 11% to 70% or a hugely regressive overall tax burden ranging from nearly 100% to 16%. The assumptions chosen were always consistent with some plausible underlying model of the economy.

To make the overall tax system look highly progressive, Whalley selected assumptions that removed the regressive taxes and made the progressive taxes more progressive. Three assumptions mattered the most, two related to the regressive taxes and one to the progressive taxes. Regarding the regressive taxes, he adopted the Browning–Johnson lifetime argument for the sales taxes that makes them slightly progressive. He also assumed that the payroll tax for Social Security was a benefits-received tax. The argument here is that the payment of the tax comes with an implied promise by the government that employees will receive a public pension during their retirement years, so that people view the tax as equivalent to a contribution to a private pension plan. This is not an implausible assumption. Many economists have argued that political support for the US Social Security system rests on just such an implied promise, even though the system was never designed to be actuarially sound, unlike private plans. Regarding the progressive taxes, Whalley noted that neither the personal nor corporate income taxes correct for inflation in calculating the tax liability on income from capital. The failure to adjust for inflation led to extremely high effective tax rates on income from capital during the 1970s and 1980s when inflation was fairly high, and thereby made each tax much more progressive in real terms, especially the corporation income tax.

To make the overall tax system highly regressive requires the reverse strategy: remove the progressive taxes and make the other taxes as regressive as possible. The way to remove the progressive taxes is to eliminate any tax burden on capital. One plausible assumption for the Canadian economy is that the return to capital is set on the world market, so that all taxes on the demanders of capital are passed on either to consumption or labor. This assumption alone sharply reduces the progressivity of the corporation income taxes and the local property taxes. Regarding the other taxes, Whalley adopts the annual perspective on sales taxes, which makes them highly regressive. He further assumes that labor income consists of a base level of income equal for all workers, augmented by income in the form of a return to a worker's accumulated human capital. Further, he assumes that physical and human capitals are perfect substitutes in production so that they must receive the same rate of return. But the return to physical capital is set in the world market and escapes any burden of taxation. Therefore, the return to human capital must also escape the burden of all taxes. This implies that all taxes on labor income are borne entirely by the base component of income. Since the ratio of the base component to total income falls sharply as total income rises, taxes on labor income are highly regressive. Notice also that the assumptions on consumption and income combine to change the corporation income and local property taxes from progressive to highly regressive, since they are now allocated either to annual consumption or base labor income. Small wonder then that the overall tax burden on low-income taxpayers approaches 100% under this set of assumptions.

Assuming perfect substitutability of physical and human capital may be extreme, but if the return to physical capital is effectively untaxed then it is reasonable to assume that at least some portion of the return to human capital is protected from taxation. If so, then the central-variant assumptions could well be wide of the mark and make the tax system seem far more progressive than it is. This conjecture is tempered by the ad hoc nature of all the assumptions in the sources and uses approach. One is hard pressed to know what to believe, which is the principal message that Whalley wanted to convey.[16]

Pure Lifetime Tax Incidence

The final stage in the evolution of the sources and uses approach was to move entirely to a lifetime perspective. Switching from an annual to a lifetime perspective has two immediate effects on the sources and uses of income.

The first is that the sources and uses are quite different in a lifetime context. The only three sources of income in a lifetime perspective are inheritances, the present value of labor income, and the present value of public and private transfer income. Income from capital drops out of the sources side, because any income from capital is assumed to grow at the same rate of return as the discount rate used to compute the present value of the income stream. The main item on the uses side is the present value of

16. Gilbert Metcalf has published an interesting analysis of pollution taxes from the sources and uses perspective. Pollution taxes are trumpeted as leading to a "double dividend" of efficiency gains. They promote efficiency directly by correcting for the pollution externality and indirectly because the revenues collected can replace revenues from other distorting taxes. The problem, however, is that pollution taxes tend to increase the prices of consumer goods, so that they are highly regressive from an annual sources and uses perspective (less so from a lifetime perspective). Metcalf considers various offsetting tax reductions to reduce the overall tax regressivity. In one experiment, he replaced 10% of personal income tax receipts with taxes on carbon emissions, gasoline consumption, air pollution, and the use of virgin materials in production. Using standard sources-and-uses data sources and methodology, he showed that the (annual) regressivity of the pollution taxes can be mostly offset if the 10% reduction in personal tax receipts is achieved by (1) removing the first $5000 of the tax base under the payroll tax, (2) offering a $150 refundable credit for each personal exemption taken under the personal income tax, and (3) cutting all personal income tax rates by 4%. He notes that cutting the corporation income tax may generate the biggest double-efficiency dividend, but that it is highly regressive from an annual sources and uses perspective. See Metcalf, 1999.

consumption, including bequests as the final act of consumption. Income from capital also appears on the uses side because saving affects the timing of consumption. Also, taxes on income from capital reduce the after-tax rate of discount that individuals use to calculate present values and thereby affect the total value of lifetime resources or consumption.

A second effect of switching to a lifetime perspective is that the variation in both the sources and uses of income across families and individuals is sharply reduced. On the sources side, the Gini coefficient for the present value of lifetime labor earnings is about half the value of the Gini coefficient for annual labor earnings. Also, because transfer receipts are concentrated among the young and the elderly, they too show much less variation over lifetimes. On the uses side, the ratio of lifetime consumption to income is approximately equal to one. The only exceptions are those families and individuals who leave substantial bequests, a very small minority. An important implication of the reduced variation in lifetime sources and uses is that selecting different incidence assumptions makes much less of a difference than in an annual context. The only way that a tax can be highly progressive or regressive is if the structure of the tax itself makes it so. Thus the overall tax system is expected to be at most only mildly progressive or regressive from a lifetime perspective.

This expectation was borne out by James Davis et al. in their 1984 study of the Canadian tax system, the first truly lifetime incidence analysis in the sources and uses tradition (Davies et al., 1984). They began by collecting a sample of 500 lifetime income profiles, spanning the full range on incomes, from the Survey of Consumer Finances. The profiles included lifetime labor earnings and transfer receipts, to which they added initial inheritances simulated from the actual pattern of mortality and bequests among Canadians. The individuals were then assumed to be life-cycle consumers with a bequest motive, choosing an optimal pattern of lifetime consumption over the years from 20 to 75, with death and the bequest occurring in the 75th year. The authors also took into account actual patterns of social mobility across income levels. The lifetime income profiles, combined with all the other assumptions, generated lifetime series for each of the 500 individuals on labor earnings, transfer receipts, and inheritances on the sources side and consumption and income from capital on the uses side. Finally, Pechman–Okner-style assumptions on the incidence of the major Canadian taxes were applied to the lifetime series to compute a central-variant lifetime incidence measure of the overall tax system.

The Canadian taxed system proved to be mildly progressive throughout: moderately progressive in the lowest four income deciles, only slightly but steadily progressive from deciles 5 through 8, and then a bit more steeply progressive over deciles 9 and 10. This was essentially the same pattern as the Pechman–Okner central variant for the US tax system and the Whalley central variant for the Canadian tax system, both from an annual perspective. As expected, the Davis et al. pattern of lifetime incidence was not very sensitive to changes in the central-variant assumptions.

A major caveat of the lifetime perspective is the assumption that individuals consume and save according to the life-cycle hypothesis (LCH). The LCH has not been supported in empirical studies of consumption and saving behavior, even when a bequest motive is added as in Davies et al. Nor has any other model of consumption and saving behavior stood up to empirical testing. Economists have not been able to reach a consensus on the best way to model consumption and saving.

In conclusion, the sources and uses approach suggests that the US (and Canadian) tax system is mildly progressive throughout, especially if one adopts a lifetime perspective. The annual incidence is more problematic in light of Whalley's caution about knowing which ad hoc incidence assumptions are the most plausible for each of the major taxes. The lifetime perspective is not without its problems, however, given the uncertainties surrounding consumption and saving behavior.

Lorenz–Gini Measures of Tax Incidence

A final development in the sources and uses tradition has been to summarize the overall effect of the tax system on the distribution of income using variations of standard Lorenz curve/Gini coefficient measures of inequality. Three popular measures are the tax concentration curve, differences in before-tax and after-tax Gini coefficients, and differences in before-tax and after-tax indexes of inequality that incorporate a social welfare function.[17]

Tax Concentration Curve

A tax concentration curve is a Lorenz style curve with the cumulative percentage of population, ordered by before-tax income, on the horizontal axis, and the cumulative percentage of the total tax burden suffered by the ordered population on the vertical axis. The tax burdens are determined by the sources and uses approach.[18]

The tax concentration curve measures the disproportionality of the tax system. Taxes are progressive, proportional, or regressive depending on whether the Gini coefficient associated with the curve is greater than zero

17. Lorenz curves, Gini coefficients, and indexes of inequality were discussed in Chapter 4. For an overview of the early work using this approach see Kiefer, 1984. See also his empirical companion piece in Kiefer, 1991.
18. A tax concentration curve could apply to a single tax or any combination of taxes including the entire tax system.

(the curve is below the diagonal), zero (the curve coincides with the diagonal), or less than zero (the curve is above the diagonal). Nanak Kakwani proposed subtracting the standard before-tax Gini coefficient from the tax concentration Gini coefficient to measure the extent of the disproportionality of the tax system (Kakwani, 1977). Using now-standard notation,

$$K = C_T - G_{BT} \tag{17.25}$$

where

K = Kakwani's extent of disproportionality index,
C_T = the Gini coefficient of the tax concentration curve,
G_{BT} = the Gini coefficient corresponding to the Lorenz curve with the cumulative percentage of population ordered by before-tax income on the horizontal axis, and the cumulative percentage of before-tax income on the vertical axis.

Change in the Before-Tax and After-Tax Gini Coefficients

A natural measure of the overall effect of the tax system on the distribution of income is the difference in the standard before-tax and after-tax Gini coefficients:

$$\text{Overall distributional effect} = G_{BT} - G_{AT} \tag{17.26}$$

where

G_{BT} is as defined above
G_{AT} = the Gini coefficient corresponding to the Lorenz curve with the cumulative percentage of population ordered by after-tax income on the horizontal axis, and the cumulative percentage of after-tax income on the vertical axis. (After-tax income refers to income minus the tax burden, not necessarily the tax payments.)

Pechman and Okner favored the proportionate version of this measure in reporting the overall distributional effect of the tax system, $(G_{BT} - G_{AT})/G_{BT}$. In his 1985 study, Pechman reported a proportionate reduction in inequality ranging from 0.8% for the most regressive variant to 2.5% for his most progressive variant (Pechman, 1985).

Change in a Before-Tax and After-Tax Social Welfare Index of Inequality

This measure makes use of indexes of inequality that incorporate society's social welfare judgments. An example is Atkinson's index of inequality discussed in Chapter 4, which is based on the equally distributed, equivalent level of income defined with reference to society's aversion to inequality. The difference in Atkinson's index using before-tax and after-tax incomes, $I_{BT} - I_{AT}$, measures the change in inequality resulting from the tax system as filtered through society's aversion to inequality. For instance, if society did not care about inequality (the utilitarian case), then the difference would be zero no matter what the pattern of tax burdens.

Recall that (1—Atkinson's index) measures the proportion of mean income that would yield the same level of social welfare as the actual distribution of income if incomes were equally distributed. Therefore, the difference in Atkinson's index before and after tax can be interpreted as an income measure of the social welfare gain from the equalizing effect of the tax system, assuming $I_{AT} < I_{BT}$.

Vertical and Horizontal Inequities

The Lorenz/Gini approach to determining the overall effect of the tax system can also be used to measure the extent of vertical and horizontal inequities. Of the two, vertical inequity is the more straightforward in the Lorenz—Gini framework.

Vertical Inequity

The problem of measuring the extent of vertical inequity can be seen with reference to the overall distributional effect $= G_{BT} - G_{AT}$. One difficulty with this measure is that any tax or tax system is likely to violate the Feldstein vertical equity principle of no reversals. That is, the ordering of the population on the basis of before-tax income may differ from the ordering of the population on the basis of after-tax income, which makes $G_{BT} - G_{AT}$ an incomplete measure of the distributional effect of a tax or tax system. The question arises as the extent of the reranking, that is, the extent of vertical inequity.

Measuring the extent of the reranking makes use of a concept called the income concentration curve, which is a Lorenz-style curve with the cumulative percentage of population, ordered by before-tax income on the horizontal axis and the cumulative percentage of after-tax income received by the ordered before-tax population on the vertical axis. The Gini coefficient associated with the income concentration curve is labeled C_Y. The extent of reranking, R, is the standard before-tax Gini coefficient minus the income concentration curve, or

$$R = G_{BT} - C_Y > 0 \tag{17.27}$$

R must be greater than zero in the presence of reranking because the income concentration curve is closer to the diagonal than the before-tax Lorenz curve. Reranking places a greater proportion of income lower down in the distribution when after-tax income is on the vertical axis rather than before-tax income, and the before-tax income is used to order the population on the horizontal axis for each curve. R can be thought of as a measure of the degree of vertical inequity in the tax system.

Horizontal Inequity

Horizontal inequity occurs when equals are treated unequally. Measuring the extent of horizontal inequity requires first defining the ideal tax base to determine whether equals have been treated unequally. J. Richard Aronson, Paul Johnson, and Peter Lambert (A/J/L) have proposed that the ideal tax base be comprehensive or Haig−Simons income adjusted for family size to account for different needs across families with the same incomes (Aronson et al., 1994; Aronson and Lambert, 1994). Following A/J/L, call the adjusted Haig−Simons income the equivalized income. Thus, two taxpayers with the same level of equivalized income before tax should pay the same tax. Any differences in their after-tax incomes are an indication of horizontal inequity, the result of inappropriate deductions, exclusions, and other "loopholes" in the tax structure.

A/J/L measure the extent of horizontal equity as follows. Define the idealized tax function, pictured in Fig. 17.3, as

$$Y_i = T(X_i) \tag{17.28}$$

where

X_i = the equivalized before-tax income of taxpayer i
Y_i = the equivalized after-tax income of taxpayer i, assuming no loopholes in the tax structure (i.e., all equivalized Haig−Simons income is subject to tax). The curve is concave because the rate structure is graduated.

Consider three levels of before-tax equivalized income—X_1, X_2, and X_3—and refer to Fig. 17.4. The presence of unwarranted tax loopholes produces a fan pattern of after-tax equivalized incomes for each before-tax level of income, as indicated in the figure. Horizontal inequity exists within each fan because equals are being treated unequally by the tax structure. Vertical inequity occurs in the regions of overlap between two fans, the regions in which taxpayers with higher equivalized incomes before tax end up with lower equivalized incomes after tax.

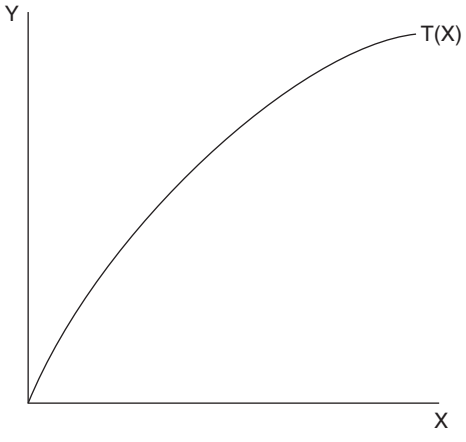

FIGURE 17.3

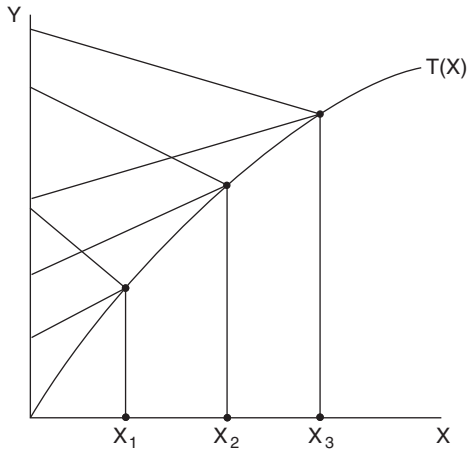

FIGURE 17.4

Aronson, Johnson, and Lambert developed the following decomposition of the overall distributional effect that includes both the horizontal and vertical inequities:

$$G_{BT} - G_{AT} = \left(G_{BT} - G_0\right) - \sum_X \alpha_x G_{Fx} - R \tag{17.29}$$

where

G_0 = the after-tax Gini coefficient that would exist if there were no fans and therefore no horizontal or vertical inequity,
G_{Fx} = the Gini coefficient within the fan associated with before-tax income X (the Gini coefficient is calculated using the after-tax order of population within the fan),
$\alpha_x = (N_x/N)(N_X\mu_x/N\mu)$,
N_x = the number of taxpayers within the fan associated with before-tax income X,
N = total population,
μ_x = the mean after-tax income within the fan associated with before-tax income X,
μ = the mean after-tax income over all taxpayers.

The second term on the right-hand side of Eqn (17.29), $\Sigma X_{\alpha x} G_{FX}$, measures the extent of horizontal inequity. It equals the sum of the income-share- and population-weighted Gini coefficients within the fans. The third term, R, is the extent of the reranking described above, the regions of overlap across the fans. It is the measure of vertical inequity.[19]

19. The decomposition is only approximately correct because the within-fan Gini coefficients are of necessity based on the after-tax ordering of the population within each fan, whereas the other Gini coefficients are based on the before-tax ordering of the overall population.

Aronson and Lambert applied their formula to the British personal income tax, using intervals of 5 Ł/week in measuring before-tax income to get enough variation in the data to produce fans. They found that horizontal inequity accounted for only about 0.2% of the overall distributional effect and vertical inequity accounted for anywhere from 4% to 6% of the overall distributional effect for the years they studied.

The A/J/L decomposition, and others like it, is subject to two caveats. The first is Feldstein's observation that a tax system cannot give rise to horizontal inequities after the market system has fully adjusted to it. Only tax reforms generate horizontal inequities, and then only temporarily. One could counter that the economy is unlikely to be in its long-run equilibrium under a given tax system. Alternatively, A/J/L's finding of very little horizontal inequity could imply that the British economy has almost completely adjusted to its tax structure, in line with Feldstein's observation. Either interpretation is a stretch, however, since the A/J/L decomposition ignores market responses to taxation and characterizes horizontal inequities only in terms of tax payments relative to their view of ideal equivalized incomes.[20]

The second caveat relates to the questionable role of horizontal equity in mainstream tax theory, raised persuasively by Louis Kaplow. As noted in Chapter 11 and elsewhere in the text, mainstream economists prefer to think of the design of tax policy in a social welfare maximizing framework, in which the value of redistributing given by the chosen social welfare function is tempered or constrained by the inefficiencies of raising tax revenue. In this framework, it is the final outcome as given by the first-order marginal conditions of the maximization that matters, the final optimal distribution, not the before and after level comparisons of horizontal equity. Horizontal equity is suspect from the social welfare perspective because of the implicit weight it gives to the original distribution, in the sense of trying to maintain some of that distribution (when combined with Feldstein's no reversals principle of vertical equity). To give an admittedly extreme example, suppose two people have equal before-tax incomes, one an honest entrepreneur and the other a thief. The social welfare framework could take this distinction into consideration, whereas the horizontal equity principle does not. Barring such examples, if the social welfare function is utilitarian in the sense that differences in the marginal social welfare weights depend only on income, then equals will be treated equally anyway once the economy reaches its long-run equilibrium. All important distinctions among individuals are matters of vertical equity in the utilitarian social welfare framework. This was essentially Richard Musgrave's point, also noted in Chapter 11, when he proposed that the horizontal equity principle should be viewed simply as ruling out inadmissible distinctions among taxpayers, such as varying tax treatment on the basis of race or gender. In summary, measures of the extent of horizontal inequity based on actual tax payments are not especially compelling from the mainstream perspective.[21]

Lorenz Measures and Tax Progressivity

The analysis of tax progressivity has long been grounded in the Lorenz tradition. We conclude this section with a brief discussion of the progressivity literature.

A tax is considered to be unambiguously distributionally progressive if the distribution of after-tax incomes weakly Lorenz dominates the before-tax distribution, that is, the distribution of average tax incomes is always at least as equal as the distribution of before-tax incomes. This notion of progressivity justifies the use of the distribution of after-tax rates (burdens) to measure progressivity, since the condition of weak Lorenz dominance is satisfied if the average tax rate is nondecreasing. Notice that proportionality is included under progressivity using weak Lorenz dominance.

Anthony Shorrocks and a number of coauthors have recently explored the question of designing income tax schedules to be distributionally progressive for all possible distributions of income. The range of choices turns out to be remarkably limited for realistic income taxes. For example,

1. In one paper, the authors show that increasing the exemptions under an income tax may not increase the progressivity of the tax if the tax has graduated marginal tax rates. It will always be progressivity increasing only under a single proportional tax rate (Keen et al., 2000). The intuition behind the result is that an increase in the exemption is worth more to taxpayers with higher marginal tax rates.
2. In another paper, the authors consider a heterogeneous taxpaying population that varies by income and need (e.g., family size). They use well-established extensions of the Lorenz dominance criterion for heterogeneous populations and analyze tax schedules that attempt to adjust for need. The federal personal income tax is an example; it allows a personal exemption based on family size and extends the tax brackets for married couples. They show that the only way to guarantee progressivity for all possible distributions is if the tax structure does not adjust for need. Further, if society

20. For a full discussion of the A/J/L approach and analysis of taxation in the Lorenz–Gini tradition, consult Lambert, 1993b. A shorter, but excellent overview of this literature by Lambert is Lambert, September 1993a.

21. Kaplow, 1989. Kaplow discusses many other potential difficulties with attempts to measure the extent of horizontal inequity in taxation.

wants to ensure that the proportion of taxes paid by the neediest group never increases, then the tax must be a proportional tax. It cannot be redistributive.[22]

The requirement that progressivity hold for all possible distributions may be overly restrictive. The authorities may know that a given tax reform is distributionally more progressive for the existing distribution. Then again, it may well be difficult to ensure that a given tax reform is uniformly progressive in the Lorenz sense throughout the entire distribution of income, if this is what the society desires.

COMPUTABLE GENERAL EQUILIBRIUM MODELS OF TAX INCIDENCE

The CGE approach to modeling an economy was made possible by Herbert Scarf's algorithm for solving complete general equilibrium models of a stylized economy, which Scarf published in 1967 (Scarf, 1967, 1969). The application of CGE techniques to overall tax incidence became popular in the 1970s, with John Shoven and John Whalley leading the way,[23] and CGE modeling is still very much in use today.

CGE models of tax incidence are the discrete version of Harberger's general equilibrium marginal analysis. Their appeal is that they can consider very broad incidence questions, such as the overall incidence of the federal and state personal income taxes or the incidence of replacing income taxes with expenditure taxes, within the context of quite detailed models of actual economies.

The typical CGE models that have been used for incidence analysis are impressively complex, even the earliest models. They contain a number of different consumers and commodities. The consumers span the full range of the income distribution, with identical consumers or a single representative consumer within each income class. The commodities are a combination of final consumer goods and intermediate inputs. The consumers' utility functions are usually specified as CES, defined over leisure and the final goods. The commodities are produced with aggregate production functions, also usually of the CES form, using labor, a given stock of capital, and a subset of the intermediate inputs.[24] The selection of inputs for each production function is guided by input/output (I/O) tables of the actual economies under investigation. The I/O tables are also used to determine the distribution of the final goods among the various income classes. The underlying market environment is assumed to be perfectly competitive.

The government sector typically consists of a Samuelsonian public good and one or more of the major taxes and transfer programs. The public good is usually not well specified. It either has no effect on consumers' utilities or it enters the utility functions in an additively separable manner so that it does not affect the various marginal rates of substitution between the commodities and leisure. The main function of the public good is to determine the resources available for private consumption. The taxes and government transfer payments, in contrast, tend to be more realistic approximations of actual taxes and transfers. The one exception is the existence of a lump-sum tax, which can be varied to consider the incidence of a single tax or transfer program in the Harberger manner. The government's budget is assumed to be balanced.

The parameters of the preference and production functions are determined in one of three ways. Some parameters are taken from existing econometric studies. Other parameters (but not all others) are set by assumption so that they can be varied as part of a sensitivity analysis of the results. The final set of parameters is residual, determined as part of a calibration exercise. Given the first two sets of parameters, the residual parameters take on whatever values are necessary such that the private and public sector variables are initially equal to their values in the actual economy being investigated. The model is said to be calibrated to an actual economy in this way.

The standard incidence exercise is to vary some combination of the public good, the taxes, and the transfers and compute the new general equilibrium. The fiscal variables are changed so as to maintain a balanced budget. The relative incidence of the tax and expenditure changes is then measured by computing HCV or Hicks' equivalent variation (HEV) for each class of consumers. For example, the HEV is the lump-sum income each consumer is willing to give up to return to the original prices (assuming a tax increase). These incidence measures typically mix actual and compensated equilibria because the consumers do not return (receive) the income lump sum required for compensation as part of the fiscal policy exercise. The new general equilibrium is the actual equilibrium, not the new compensated equilibrium. Thus, the incidence measure can be thought of as focusing on each individual income class one at a time, under the implicit assumption that if only one consumer were compensated lump sum it would have no significant effect on the new general equilibrium. This is not quite accurate, as all consumers are affected simultaneously by the fiscal experiment. Nonetheless, the HEV measures give a sense of the relative burdens suffered by each income class, which is the purpose of the exercise.

22. Moyes and Shorrocks, 1998. Interested readers should also see Ebert and Moyes, 2000.

23. For an overview of these models and the related literature, see Shoven and Whalley, 1984.

24. The first CGE models were one-period static models, without saving and investment decisions or capital markets. The extension of these models to a lifetime context with some dynamics appeared in the 1990s. The lifetime CGE modeling approach is discussed in the final section of this chapter.

The big advantage of the CGE approach to tax incidence relative to the sources and uses approach is that it can approximate the dead-weight losses of the distorting taxes and transfers as wages and prices change from one general equilibrium to another. It does not compute them exactly because the new general equilibrium is not the compensated equilibrium. Even so, having a sense of the relative dead-weight losses is important because each individual's dead-weight loss is the proper measure of incidence or burden in a compensation experiment.

Despite their very different approaches, the CGE models and the sources and uses approach have reached roughly the same conclusions about the overall incidence of the five major taxes in the United States. They agree that the overall US tax system is mildly progressive. The agreement of the two approaches is convenient because otherwise researchers and policy makers would be forced into making a choice in the nature of the lesser of two evils. Are they willing to accept the heuristic and sometimes problematic incidence assumptions of the sources and uses approach? Alternatively, are they willing to accept the many assumptions required to specify and parameterize the preferences, production functions, and public sector variables in the CGE models, as well as the assumption of perfectly competitive markets? Which set of assumptions to prefer is unclear, especially since the CGE models are highly simplistic representations of actual economies despite their mathematical complexity.

DYNAMIC TAX INCIDENCE

The early models of tax incidence were static, one-period models. Dynamic models were sure to follow, however, because the dynamic analysis of tax incidence has three huge advantages over static analysis.

First and foremost, a dynamic model can track the evolution of the capital stock in response to tax policies. Changes in the capital stock through time affect the marginal products of capital, labor, and all other factors of production, which determine the real returns to the factors in a competitive environment. These capital-induced changes in the returns to factors over time tend to swamp any direct short-run effects that tax policies might have on factor returns.

Second, a dynamic model can consider intergenerational tax and expenditure incidence, that is, the relative effects of government policies on people of different ages, such as the young who are still working and the elderly who are retired. Dynamic models have shown that these intergenerational effects can be very large and also very important to the evolution of the economy.

The third advantage, related to the second, is that a dynamic model can analyze the incidence of the asset revaluations that immediately follow changes in government policies. The asset revaluations occur because capital is costly to adjust, so that capital assets of different vintages are not perfect substitutes in production. In fact, the short-run supply elasticities of many kinds of capital are quite low. Therefore, as changes in tax policy change the demands for different kinds of capital, fairly large changes in capital prices may be required in the short run to maintain equilibrium in the capital markets. An example is an investment tax credit, which favors new capital over existing ("old") capital and thereby lowers the relative price of new versus old capital. These asset revaluations matter in a dynamic context because people of different ages tend to hold different proportions of old and new capital. For example, the retired elderly have a much higher proportion of claims to old capital in their portfolios than do the working young. Dynamic tax analysis has suggested that the most important incidence effect of tax policy in the short run may well be the intergenerational transfers of wealth through asset revaluations following the change in taxes. (We will return to this point below.)

The two growth models most commonly used in dynamic incidence analysis are the Ramsey model with an infinitely lived representative consumer and the overlapping generations (OLG) model with two or more cohorts of finitely lived consumers. The Ramsey model appeared first, but it was soon overtaken by the OLG model as the preferred model for tax incidence. Only the latter can analyze intergenerational incidence and, as indicated, the early OLG models showed just how important the intergenerational effects of tax and expenditure policies can be.

Peter Diamond's (1965) article comparing and contrasting the burdens of internal and external public debt was the seminal application of the OLG model to a fiscal policy issue (Diamond, 1965). Once Diamond had demonstrated the advantages of the OLG framework, other economists were quick to apply it to other fiscal issues, including tax and expenditure incidence.[25] Foremost among them were Alan Auerbach and Lawrence Kotlikoff, who subsequently became the economists most closely identified with OLG incidence analysis. They published their complete OLG model with applications to a number of fiscal policy issues in *Dynamic Fiscal Policy* (Auerbach and Kotlikoff, 1987; Kotlikoff, 1984; Kotlikoff and Summers, 1985).

The basic Auerbach–Kotlikoff model and its variations are far more complex than Diamond's original model, so much so that we will only provide a sketch of the main features of the model. Our goal is simply to give a sense of the various kinds of incidence channels that can occur in an OLG framework. Even the simplest Auerbach–Kotlikoff-style OLG models are so complex that they must be solved with simulation techniques.

25. The OLG model also quickly gained favor with macroeconomists for analyzing long-run macroeconomic issues.

The Auerbach–Kotlikoff OLG Model

Structure of the Model

The baseline Auerbach–Kotlikoff model has five essential elements that determine how the economy evolves over time: production, consumption, the government sector, the underlying market environment, and assumptions about how people form expectations of the future and what they know at any time.

Production

The production side of the model is highly simplified. A single all-purpose good, Y, is produced each period using capital and labor. The aggregate production function is CES. Y is either purchased by consumers as their one consumption good or purchased by the government, in which case it becomes a Samuelsonian nonexclusive public good.

The cost of adjusting the capital stock each period is a function of the level of investment and the investment/capital ratio, such that the marginal cost of investment is a linear function of the investment/capital ratio. The investment decision follows Tobin's q theory of investment with costly adjustment of capital. The marginal cost of investment incorporates tax variables such as a tax on capital income and an investment tax credit.

Consumption

Consumers make economic decisions for 55 years, from ages 21 through 75 years, after which they die. The model is simulated for 155 years (represented by ∞ below), so that a large number of cohorts (generations) are alive at any one time. Consumers maximize lifetime utility according to the LCH, with utility a function of their consumption and leisure during each period of their economic lives. (There are no bequests in the baseline model.) The utility function is CES within each period and additively separable over time, discounted by a rate of time preference that is the same for everyone. Consumers are endowed with one unit of time each period that they allocate between labor and leisure. The amount of leisure taken each period and the time of retirement are endogenous.

One of the three main equations of the model that drive its results is each consumer's intertemporal budget constraint. Assuming no borrowing or lending constraints, the most basic intertemporal budget constraint for a consumer of cohort j, one without taxes, is

$$\sum_{t=0}^{T}\left(\frac{C_{jt}}{(1+r_t)}-\frac{we_t(1-1_t)}{(1+r_t)}\right)=0 \qquad (17.30)$$

where

0, T = the initial and final periods of consumer j's economic life,
C_{jt} = consumption by consumer j and period t,
r_1 = the t-period interest rate for money received in period t; $(1+r_t)$ equals $\prod_{s=0}^{t}(1+r_s)$,
where r_s is the one-period interest rate in period s,
w = the wage for an unskilled unit of labor,
e_t = a skill parameter that permits an age–earnings profile to be built into the model,
l_t = the amount of leisure taken in period t; $l_t = 1$ indicates retirement.

The Government Sector

The government exogenously sets the value of the public good and the structure of taxes and transfers each period subject to an intertemporal budget constraint that satisfies the no-Ponzi condition. The public debt must be bounded; debt cannot grow indefinitely at a rate above the rate of interest (which is greater than the growth of the economy in the long run in the Auerbach–Kotlikoff model). The government can run a deficit in any one period, however, which it covers by issuing bonds that mature after one period. These assumptions lead to a single-period government budget constraint of the form (assuming no money)

$$D_{t+1}-D_t = G_t + r_tD_t - T_t \qquad (17.31)$$

where

D_t = debt issued in period t
G_t = the public good in period t
r_t = the government's one-period borrowing rate
T_t = taxes–transfers in period t

Adding the single-period budget constraint over all periods yields the government's intertemporal budget constraint under the no-Ponzi condition:

$$\sum_{t=0}^{\infty}\left(\frac{T_t-G_t}{(1+r_t)}\right)=D_0 \qquad (17.32)$$

where r_t once again refers to the t-period interest rate for money received in period t. The government's intertemporal budget constraint is the second of the three main equations that drive the results.

The third main equation is the aggregate intertemporal consumption possibilities for the economy.

$$\sum_{j=1}^{J}\sum_{t=0}^{\infty}\frac{Z_{jt}}{(1+r_t)}+\sum_{t=0}^{\infty}\frac{G_t}{(1+r_t)}=H_{\mathrm{PV}}+A_0 \qquad (17.33)$$

where

Z_{jt} = consumption of goods and leisure by person j in period t,

H_{PV} = the present value of the aggregate lifetime labor endowment, discounted to time zero (the endowment includes the time available for labor or leisure),
A_0 = the economy's initial endowment of physical capital to be used in production.

All resources are ultimately used to produce the consumption good or the public good.

The Market Environment

The economy is assumed to be perfectly competitive. As noted earlier, this implies that the real returns to labor and capital equal their marginal products. Further, labor and capital are fully employed each period, even while the economy is in transition moving from one steady state to another steady state following a change in government policy. The transition to the new steady state takes 150 periods.

Consumers' Expectations and Their Information Set

The guesses consumers and producers make about the future values of all the variables that are exogenous to them—prices, endowments, and tax (transfer) rates—determine their behavior currently and in all future periods and thus the general equilibrium in the economy period by period. This is why an assumption about how people form expectations of the future is a crucial element of any dynamic model. Auerbach and Kotlikoff assume that people have perfect foresight regarding these variables, meaning that their predictions lead to general equilibria each period that generate precisely the values they predicted.

People are not omniscient in the Auerbach–Kotlikoff model, however. Changes in the government's policies catch them by surprise as they occur and cause them to reoptimize from that time forward, again with perfect foresight about prices, endowments, and tax (transfer) rates. Since utility is additively separable over time, all past behavior is irrelevant to their reoptimizations. Also, consumers are partially myopic. They do not see the aggregate consumption possibilities or the government's intertemporal budget constraint embedded within it. Hence, they do not understand how the government will adjust to their reactions to fiscal policies. These informational assumptions are crucial because if people did know the aggregate consumption possibilities and the government's intertemporal budget constraint, then fiscal policies would not have any effect. Having once determined an optimal lifetime path of consumption and leisure and knowing the time path of G, consumers would know how to reoptimize from that time forward in such a way as to offset fully anything the government might try to do with its tax (and transfer) policies. Having internalized the government's responses to their decisions, all taxes would become in effect lump sum taxes that consumers could offset.

Fiscal Policy Options

Given the structure of the model, fiscal policies can have real effects on the economy only by changing the consumers' intertemporal budget constraints and thereby altering consumer behavior. Their budget constraints depend upon the net-of-tax prices, wages, and interest rates; labor earnings; capital endowments; and lump-sum taxes and transfers. With this in mind, fiscal policy can essentially do four things in an OLG framework:

1. Change marginal incentives, the net-of-tax prices, wages, and interest rates.
2. Increase or decrease spending on the public good G, which changes the aggregate endowments available for consumption through the aggregate intertemporal consumption possibilities frontier.
3. Redistribute resources across generations (intertemporal redistribution).
4. Redistribute resources within generations (intratemporal redistribution).

Some comments on the first three options are in order.

Changing Marginal Incentives

A change in net-of-taxes prices, wages, and interest rates affects consumers in three ways. It changes the relative prices of present and future consumption and leisure that consumers equate to the marginal rates of substitution between these variables. It changes the present value of labor and capital endowments by changing the net-of-tax wages and interest rates. And, finally, it changes the incentive to invest in human capital and therefore affects the evolution of future wage rates.

Changes in the Public Good

The treatment of the public good in the Auerbach–Kotlikoff model is similar to that in the CGE models. G does not generate utility; its only effect is to increase or decrease the resources available for consumption. G could be modeled to have a direct effect on utility without changing the general implications of the model. The only requirement then would be that G not be a perfect substitute for consumption. If it were, it essentially disappears from the aggregate intertemporal consumption possibilities relationship. Public goods are unlikely to be perfect substitutes for private goods, however.

Intertemporal Redistributions

Intertemporal redistributions can have major effects in OLG models with LCH consumers, especially if there is no bequest motive. The reason is that consumers react to one-time changes in their endowments by spreading the higher

or lower consumption possibilities over their remaining lives. Older consumers naturally have higher marginal propensities to consume out of changes in endowments than do younger consumers, simply because their consumption is spread over fewer remaining years. The differences in their MPCs imply that transfers from younger to older generations increase aggregate consumption and reduce saving and investment. Conversely, transfers from older to younger generations decrease aggregate consumption and increase saving and investment. The natural increase in MPCs as cohorts age is a central feature of the OLG model.

With these thoughts in mind, Auerbach and Kotlikoff describe four specific kinds of dynamic fiscal policies:

1. Tax substitutions—In a pure tax substitution, one tax is reduced and another tax increased such that there is no change in total tax receipts at the time of the tax substitution or for any period thereafter. The new tax always raises the same amount of revenue that the old tax would have raised. The public good also remains unchanged over time. Although tax substitutions involve no aggregate redistribution from the private to the public sector, they do have income effects as well as substitution effects in an OLG framework because different cohorts are affected differently. For example, the substitution of a wage (payroll) tax by a consumption (expenditures) tax disproportionately burdens the elderly, because a wage tax is paid only until retirement, whereas a consumption tax is paid until death.
2. Balanced budget changes in expenditures and taxes—A one-time change in G is matched by a change in taxes or transfers each period such that $(G_t - T_t)$ remains constant, equal to its value before the change in G.
3. Temporary deficits or surpluses (intergenerational redistributions)—A pure temporary deficit consists of a decrease in a particular tax (increase in a transfer) for a number of periods, followed eventually by an increase in that same tax (decrease in the same transfer) such that the accumulated debt per person during the temporary period remains constant forever after. Permanent deficits are not allowed because they would violate the no-Ponzi condition on the government's intertemporal budget constraint. A temporary surplus is the reverse of a temporary deficit.

In an OLG framework, temporary government deficits and surpluses are defined as intergenerational redistributions. A temporary deficit is any fiscal policy that causes a redistribution from the younger to the older generations. In the case of a temporary tax cut, the older generations gain more from the temporary reduction in their taxes than they lose later on when the taxes are increased. The reverse is true for the younger generations. Furthermore, since the older generations have higher MPCs than the younger generations, a temporary deficit increases aggregate consumption, thereby reducing aggregate saving, investment, the future stock of capital, and the productivity of the economy in the long run. Running large deficits has been the policy stance of the federal government since the 2001 tax cuts under George W. Bush and the deficits are expected to continue indefinitely without expenditure cuts or tax increases. They represent a substantial drag on long-run economic growth.

Conversely, a temporary surplus is any fiscal policy that causes a redistribution from the older to the younger generations. In the case of a temporary tax increase, the older generations lose more from the temporary increase in their taxes than they gain later on when the taxes are reduced. The reverse is true for the younger generations. Consequently, a temporary surplus such as the United States experienced briefly at the end of the 1990s decreases aggregate consumption, thereby increasing saving, investment, the future stock of capital, and the productivity of the economy in the long run.

The analysis of temporary surpluses and deficits has three important implications for fiscal policy in the long run:

1. Annual deficit/surplus measures—Measures of annual government budget deficits and surpluses are irrelevant in a dynamic framework. The productivity of the economy in the long run is determined entirely by the amount of spending on the public goods and the extent of the redistributions across generations.
2. Ricardian equivalence—The potentially large impact of intergenerational redistributions on productivity arises only because the young and the old have different marginal propensities to consume, as they surely would without a bequest motive. Robert Barro has proposed an alternative model with bequests that removes these differences (Barro, 1974). He assumes that the older generations are altruistic toward the younger generations and will not allow them to be affected by temporary deficits and surpluses. For example, instead of consuming all of a temporary decrease in taxes over their remaining lifetimes, the elderly save just enough of the decrease to pass on a bequest to the younger generations that removes the relative disadvantage the young would otherwise suffer. With the relative burdens across generations equalized through the bequests, the temporary deficit has no effect. The same argument holds in reverse for temporary surpluses: The elderly reduce their bequests to remove the relative burden they themselves would otherwise suffer under a temporary surplus. This ineffectiveness of intertemporal redistributions is known as Ricardian equivalence. Barro's OLG model with

altruism and bequests implies that the amount of the public good (i.e., public consumption or investment) is the only fiscal policy variable that has a real endowment effect on the economy in the long run. Although Ricardian equivalence is possible, it is extremely unlikely to hold in practice. Most public sector economists believe that intertemporal redistributions have real and important effects on the productivity of the economy in the long run.

3. Investment versus savings incentives—The analysis of temporary deficits and surplus leads to a related distinction between investment and savings incentives. The distinction matters under the realistic assumption that adjusting the stock of capital is costly, so that different vintages of capital are not perfect substitutes in production.

   Investment incentives are policies that favor new capital over existing capital, such as accelerated depreciation allowances and an investment tax credit. These types of incentives act as hidden temporary surpluses because the older generations hold a disproportionate share of the claims to the existing capital stock. Thus, investment incentives redistribute resources from the older to the younger generations, thereby reducing aggregate consumption and increasing the productivity of the economy in the long run. A tax on land also acts as an investment incentive because land is held disproportionately by the older generations. Finally, a tax substitution of replacing an income tax with a consumption tax can be thought of as a self-financing investment incentive since the older retired generations lose under the substitution. They have already paid income taxes while working, and they will now pay taxes again on their consumption during retirement.

   Savings incentives are policies that favor new and old capital equally, such as lower taxes on capital gains, tax exemptions on municipal bonds, and tax-deferred pension funds and individual retirement accounts (IRAs). A tax substitution of replacing an income tax with a wage (payroll) tax can be thought of as a self-financing savings incentive since the return to all capital becomes untaxed under this substitution. Because savings incentives favor both kinds of capital equally, they are less potent stimulants to saving and investment per dollar than are investment incentives. They do not hit the older generations as hard.

4. Intratemporal redistributions—These are balanced-budget redistributions within each generation, such as an annual redistribution from the nonpoor to the poor. Intratemporal redistributions are the only fiscal policies that do not have special effects in a dynamic OLG setting because they do not transfer resources across generations.

A Selection of Results

Realistic fiscal policy changes tend to have fairly potent long-run effects in the Auerbach—Kotlikoff model. Here are some examples:

Tax Substitutions

Auerbach and Kotlikoff consider the replacement of a 15% income tax with a consumption tax, a wage tax, and a capital income tax. The first substitution raises average welfare each year in the new steady state, and the last two lower average welfare[26]:

1. Consumption tax—Replacing an income tax with a consumption tax sharply favors the young and future generations over the older generations. As such, it leads to large increases in saving, investment, capital stock, productivity, and wages before tax. Overall welfare increases by 2.22% per year in the new steady state.
2. Wage tax—In this substitution, the elderly gain relative to the young and future generations. Nonetheless, saving, investment, capital stock, and productivity increase because returns to capital are no longer taxed. Overall welfare declines by 0.89% per year in the new steady state, however, primarily because after-tax wages decrease. The wage tax does not enhance productivity as much as the consumption tax because it favors rather than hurts the elderly. In an OLG context, a consumption tax is equivalent to a wage tax plus a one-time levy on existing capital of the elderly. The levy equals the present value of the taxes the elderly would have paid under a consumption tax.
3. Capital income tax—This substitution lowers overall welfare by 1.14% per year in the new steady state. By increasing the burden on future consumption over current consumption relative to the income tax, a tax on capital income lowers saving, investment, the capital stock, and productivity.

26. The ultimate interest in the policy simulations is the change in consumers' welfare, which can be handled in a number of different ways in a dynamic context. Suppose a policy generates productivity gains and increased national output. Auerbach and Kotlikoff choose a dynamic HEV measure constructed as follows. First, they introduce a distributional authority that continually redistributes income lump sum to all cohorts born before a certain date so that their utility remains constant, unaffected by the policy change. It then distributes all the remaining gains lump sum and equally to the cohorts born after that date. Having redistributed all the increased output, they measure the gain in welfare as the HEV for the latter cohorts, equal to the lump-sum income each cohort would require at the original prices before the policy change to be as well off as they are following the policy change. This method of compensating the two sets of cohorts from the productivity gains tends to reduce the changes in welfare resulting from the policy simulations. The tax substitution results are taken from Kotlikoff, 1984, pp. 1601—1603.

Balanced Budget Increases in the Public Good

The principal conclusion of the Auerbach–Kotlikoff model is that the tax used to finance an increase in G matters quite a bit. For example, financing an increase in G with a consumption tax leads to a big decrease in consumption. Financing the increase with a wage tax mostly decreases the supply of labor; the decrease in consumption is much smaller.

The one drawback of this particular exercise is that G has no direct effect on utility. Therefore, the model cannot consider the possibility that a tax-financed increase in a public good may be welfare enhancing no matter which tax is used. A true balanced-budget incidence experiment would require a complete specification of the public good, including its effects on consumers' utilities, and the use of the incidence measure for a Samuelsonian nonexclusive public good discussed in the first part of the chapter.

Temporary Deficits

Perhaps the most important insight of the Auerbach–Kotlikoff model is the huge real effect of temporary deficits and surpluses. In one exercise, they reduce the income tax from 15% to 10% for 20 years. This requires an income tax rate of approximately 30% in the 21st year to hold the additional debt per person accumulated in years 1–20 constant from then on. As a result, the capital stock falls by 49%, the before-tax wage falls by 14%, and the interest rate increases by 4 percentage points in the new steady state. Keep in mind, however, that the cost of a temporary deficit is not simply that taxes eventually have to be increased, as is often alleged. Rather, it is that a temporary deficit combined with the subsequent tax increase redistributes purchasing power from the younger to the older generations, thereby increasing aggregate consumption.

Auerbach and Kotlikoff also find that asset revaluations following a change in tax policy can be very large. For example, the 1981 Tax Reform Act introduced a number of new investment incentives into the federal corporation income tax, most notably an investment tax credit on equipment and more accelerated depreciation allowances on different classes of assets. Auerbach and Kotlikoff estimated that the resulting asset revaluations led to a $260 billion loss in the value of existing capital. These tax reforms thus represented a huge stimulus to life-cycle saving by redistributing purchasing power from the older to the younger generations (Kotlikoff, 1984).

Intratemporal Redistributions

Auerbach and Kotlikoff find that intratemporal redistributions tend to have fairly modest effects. This is true even if the poor are liquidity constrained and immediately spend whatever transfers they receive each period. There appears to be an approximate balancing each year of the poor's extremely high MPC of their (relatively) small transfers and the nonpoor's quite low MPC of the (relatively) large present value of their taxes required to pay for the annual transfers.

The fiscal policy experiments with the OLG model point to an extremely unsettling trade-off between efficiency and equity in the long run. The largest gains in productivity are achieved by policies that exploit the differences in the MPCs of the younger and older generations. In other words, much of the gains in overall efficiency come at the expense of the older generation. The converse is also true. Policies designed to favor the elderly in the name of equity are likely to be harmful to productivity and overall welfare. Finally, if society simply chooses not to hurt its elderly and designs its fiscal policies to be fairly neutral toward them, then it may not be able to realize much gain in productivity or average welfare.

Concluding Caveats

Although the Auerbach–Kotlikoff OLG model has yielded many new insights about the incidence of tax and expenditure incidence relative to the older static models of incidence, the model is not without its problems. For one, the informational assumptions behind the model are highly suspect: Consumers have perfect foresight about future prices but at the same time are fooled by changes in government policy and are myopic regarding aggregate consumption possibilities. Equally problematic is the assumption of perfect competition in labor markets, that labor is always fully employed in the long transition to the steady state. Third, economists do not really understand saving behavior. The LCH of consumption and saving is a natural choice for a long-run policy model, with or without bequests, but the LCH has not been supported by empirical research. Fourth, the empirical analysis of investment has not yielded a consensus on the determinants of investment demand. Finally, the introduction of human capital into these models can have fairly dramatic effects. Some brief observations on the last three points follow.

Saving

The chief competitor to the LCH is precautionary saving, which is a response to uncertain income streams. Precautionary saving is known to be less sensitive to changes in after-tax rates of return than life-cycle saving. Eric Engen and William Gale introduced uncertain incomes and precautionary saving into a long-run OLG model and found that replacing the US personal income tax with a flat-rate consumption tax would increase saving by only 1/2%,

and steady-state gross domestic product by only 1−2%. The elasticity of saving with respect to after-tax income in their model is 0.39. When they remove the income uncertainty and the precautionary motive for saving, the elasticity rises to 1.94.[27]

Investment

The majority of studies find that the response of investment to changes in the cost of capital are quite low and operate with fairly long lags. Austan Goolsbee has suggested that these results may be due to a low supply response of many kinds of capital goods. He finds that a 10% investment tax credit on equipment raises equipment prices almost immediately by 3.5−7%, and that the price effects last for a couple of years.[28] Therefore, much of the tax incentive to stimulate investment demand is captured at first by the suppliers of capital as increased rents. The short-run price response is important, because from 1959 to 1988 Congress changed the tax laws affecting the cost of capital at least once every 4 years. Goolsbee believes that the price response explains why capital goods suppliers lobby for tax reductions, lags are important in investment demand equations, and the estimated demand elasticities of capital are so low. By taking into account the price responses, he obtains very high estimates of the demand elasticity, in the range of 0.95−2.15. Still, the response of the capital stock to tax changes is sharply reduced in the short and medium run by the low supply elasticities.

Human Capital

Human capital is a very important resource. Economists frequently cite the estimate by Davies and Whalley that the stock of human capital is three times the stock of physical capital in the United States (Davies and Whalley, 1989). Moreover, incorporating human capital into an OLG growth model can dramatically alter the effects of different kinds of tax policies.

Human capital has a number of distinctive features relative to physical capital. First, it depreciates completely at retirement, so that it cannot be passed on to heirs. Second, the taxation of human capital arises primarily from taxes on wages and salaries that reduce the returns to human capital. Whether it is heavily or lightly taxed under an income tax depends in part on how human capital is acquired. It is potentially vulnerable to heavier taxation than physical capital because there is no depreciation allowance for human capital to be taken against wage income. Thus, both the principal and returns to human capital are taxed. But, if human capital is received through an on-the-job training program and "paid for" by a reduction in wages during the training period, then the investment in human capital may be effectively expensed, in which case there is no marginal tax on human capital. Much of human capital is received through formal education, however, which is only partially subsidized. Therefore, human capital is taxed under an income tax, and potentially quite heavily. Consider the relative effects of a comprehensive income tax on physical and human capital. The taxation of interest income and other returns to capital is borne directly by physical capital but not human capital. Conversely, the taxation of wage and salary income is borne directly by human capital but not physical capital. On net, a comprehensive income tax discriminates against human capital in favor of physical capital because the majority of human capital is not expensed so that the remaining principal is taxed as well as the returns. Finally, although human capital cannot be bequeathed directly, it does contribute to the overall stock of knowledge, which has a lasting and important effect on the productivity of the economy.

Marc Nerlove et al. added human capital to Diamond's original OLG model and compared the zero-tax steady state with the steady state under various tax policies.[29] None of the human capital was expensed in their model. Among their findings are the following:

1. A proportional 50% comprehensive income tax led to a big increase in the ratio of physical to human capital and reduced productivity by 90%, but the tax increased welfare because it moved the economy closer to the Golden Rule steady state.
2. Human capital has such an important effect on productivity in their model that a wage tax lowers productivity because of its bias against human capital and a tax on capital income raises productivity because of its bias in favor of human capital. Human capital does not escape a tax burden under a capital income tax because the reduction in the stock of physical capital reduces wages. Nonetheless, the relative bias in favor of human capital under the capital tax is sufficient to increase productivity in their model.

Despite the modeling uncertainties, the OLG framework is useful for comparing the potential long-run effects of different kinds of fiscal policies. Also, its analysis of intergenerational redistributions is an important contribution to the public sector literature.

27. Engen and Gale, 1997. Part of the small saving response is also explained by the exemption from tax, or the reduced taxation, of many forms of saving under the personal income tax, such as saving for retirement, imputed rents, and capital gains.

28. Goolsbee, 1998. The supply response is especially low for equipment that has large back orders at the time of a tax change or that faces low competition from imports. For an overview of the empirical literature on investment demand, see Chirinko, 1993; Chirinko et al., 1999.

29. Nerlove et al., 1993. A more recent analysis of the taxation of human capital using a Ramsey growth model is Trostel, 1993.

The Fullerton–Rogers Lifetime CGE Model

The last stage in the evolution of tax incidence models came in 1993. Don Fullerton and Diane Rogers developed a lifetime CGE model for a study of the US tax system produced for the Brookings Institution (Fullerton and Rogers, 1993). They sought to combine the strengths of the three main models of economy-wide tax incidence at the time: the complex representation of the economy made possible by the CGE model, the lifetime perspective and some of the dynamics of the OLG model, and the use of detailed data on families and individuals that characterizes the sources and uses approach. Their model is far too complex to discuss in any detail. We will only highlight some of its novel features and the main incidence findings.

The principal innovations in their model were on the consumption side. Fullerton and Rogers began with a sample of 838 individuals from the Panel Study of Income Dynamics, which had collected economic, social, and demographic data on families and individuals for 18 years by the time of their study. They used the sample to estimate a wage equation as a function of age and other demographic variables and then combined the estimates with actual data to construct lifetime wage profiles for everyone in the sample. The lifetime wage profiles are then used to compute the present value of the labor endowment for each of the individuals in the sample, their potential lifetime incomes. The labor endowment equals the sum of the estimated wages at each age of their economic lives times 4000 h, the number of hours assumed to be available for labor or leisure each year, discounted to present value. The people are divided into 12 income classes on the basis of their potential incomes. Thus, everyone in the sample is characterized by age and income class. These constitute the consumers in their CGE model.

The other elements of the model are highly complex but in line with existing static and dynamic CGE models in the early 1990s. There are 17 different consumer goods, 5 different classes of assets (equipment, structures, land, inventories, and intangibles), 37 representative firms divided into corporate and noncorporate sectors, and production functions for each good whose arguments are capital, labor, and intermediate inputs. The choice of variables in the production functions was guided by I/O tables for the US economy. The firms' behavior is not dynamic in the sense that they make no investment decisions. The dynamics in the model center on the consumers' saving decisions, with investment set equal to saving each period.

The government produces one of the 17 consumer goods for sale to consumers, and a public good that it supplies free of charge. The public good enters separably into consumers' utility functions. The government also gives lump-sum transfer payments to consumers, which are estimated for each consumer based on the actual pattern of government transfers by age and income class. Government revenues are collected from the five major US taxes, and the government's budget is balanced each period.

Markets are assumed to be perfectly competitive. A foreign sector is added to close the model with exports equal to imports, and the model is calibrated to the US economy.

The individuals are LCH consumers with a bequest motive included, with 60-year economic lives from ages 20 to 79 years. The exogenous variables for each person are an inheritance, which Fullerton and Rogers estimate from actual data on inheritances by income class; the lifetime wage profile; a set of tax rules; and a lifetime profile of lump-sum transfer payments received from the government. Given these variables, which they know with perfect foresight, consumers maximize their lifetime utilities in a three-stage sequence. First, they decide how much of their potential income to "spend" each period on consumption and leisure versus how much to save, in accordance with a CES utility function. Next, they decide on the allocation of their "spending" between leisure time and consumer goods each period, again in accordance with a CES utility function. Finally, they determine their purchases of the 17 consumer goods each period so as to maximize a Stone–Geary utility function over the goods. The model also allocates the consumers' purchases of private goods to each of the industries, and then to the corporate and noncorporate sectors within each industry.

The tax incidence simulations are pure tax substitutions. They consist of changing one or more of the existing taxes and replacing the revenue gained (lost) with a proportionate tax cut (tax) on the consumers' potential income. Because the revenue-compensating tax is levied on potential income, it is a lump-sum tax. The CGE model simulates the effects of the tax change on the evolution of the supply of labor, saving, capital, outputs, and prices. The incidence comparisons are based on the lifetime consequences of the new evolution of the economy for the individuals within each income class. The measure of the change in economic welfare is the consumers' lifetime HEV, which, for a tax cut, is the present value of the lump-sum income that the consumer would be willing to accept to forego the price changes. The relative burdens are computed for each income class. A tax is progressive if the HEV is proportionately higher for the higher income classes. Finally, since the incidence of each of the five major taxes is determined by removing it and replacing it by the proportionate tax on potential income, the incidence of each tax is being compared relative to the incidence of a proportionate, equal-yield, lump-sum tax.

Only some of Fullerton and Rogers' results were consistent with those of the static CGE and sources and uses models at the time. In line with existing studies, Fullerton and Rogers found that the personal income tax

was progressive, with the burden in terms of lifetime potential income ranging from 5% for the lowest income class to 19% for the highest income class. They also found that the sales and payroll taxes were regressive.

The surprises were the property tax and the corporation income tax. The incidence of the property tax fell on all capital in their model because of the competitive assumptions, which tended to increase the burden on the higher income classes. But it also increased housing costs, which disproportionately burdened the lower income classes. Thus, the property tax had a U-shaped pattern of lifetime incidence. The corporation income tax was so small that it had no noticeable effect on the sources side of the model. It did, however, lead to a reallocation of production toward items that used lightly taxed forms of capital, with some resulting adjustments in the relative costs and prices of the consumer goods. The relative price adjustments turned out to be antipoor, so that the corporation income tax was slightly regressive overall.

Overall, the Fullerton–Rogers results appear to be roughly consistent with the general consensus that the overall US tax system is mildly progressive even in a lifetime context, largely because the federal and state personal income taxes are progressive.[30]

APPENDIX

Tax Reform and Tax Theory

Tax reform is a topic of unending political debate in all the industrialized market economies, the perennial questions being what is the best tax base and how progressive should the chosen tax be? The Appendix discusses what mainstream tax theory has to offer in answering these two questions, drawing on results from previous chapters and adding some considerations that are not covered elsewhere in the text.

As it happens, most economists are fairly guarded about offering advice on tax reform. The reason is that tax theory does not provide specific answers to either question. The only consensus that has emerged among economists is how to think about these questions.

The modern view follows Mirrlees approach to optimal income taxation that the government's objective should be to maximize a Bergson–Samuelson social welfare function subject to a government budget constraint that incorporates the desired tax or tax system. Economists have moved away from the Smith–Mill ability-to-pay framework with its emphasis on Haig–Simons income, even though the public may not yet have abandoned the older view. In addition, although Mirrlees' analysis of the optimal income tax was static, the appropriate objective function for considering tax reform is social welfare defined over individuals' lifetimes. Consequently, the two most common general equilibrium models in which the government's problem is embedded are the Ramsey infinite horizon representative consumer model and the OLG model with cohorts of working and retired people.

The consensus on how to think about tax reform has not, however, led to anything approaching a consensus on how to answer the two fundamental questions of tax reform. Consider the question of how progressive the tax or tax system should be. The question runs immediately into all the problems associated with the social welfare function described in Chapter 4: What is it? What should it be? Can it be?—Arrow's general impossibility theorem. We will not revisit those issues here. Suffice it to say that virtually all economic analysis of tax reform assumes that the tax system itself should redistribute from high-income to low-income individuals. This is a natural consequence of adopting the social welfare maximizing framework, since any standard social welfare function exhibits aversion to inequality.[1] And there does seem to be a consensus within the United States that the tax system should be at least mildly progressive.

Reforming the Tax Base: The Contenders

Regarding the question of the best tax base, four reform proposals for US federal taxation have received the most attention among economists. They would replace the existing US personal income and corporation income taxes in a revenue-neutral manner.

- A broad-based personal income tax with no deductions and exclusions. Some proposals retain the personal exemptions, others do not.
- A personal consumption or expenditures tax, again with and without personal exemptions
- A consumption tax in the form of a national retail sales tax, a reform most closely associated with Laurence

30. Altig et al. published an expanded version of the Auerbach–Kotlikoff model that includes many features of the Fullerton–Rogers lifetime CGE model. Their hybrid model has 12 income classes, but not as many goods and factors as the Fullerton–Rogers model. The multiple income classes allow them to analyze the effect of various tax substitutions by cohort and income class. For instance, they found that substituting an ideal expenditures tax for the current US federal personal income tax hurts the high-income elderly proportionately more than the low-income elderly. A noteworthy prediction of their model is that switching to an expenditures tax would raise steady state output in the United States by 9%. Altig et al., 2001.

1. When studies employ the Bethamite utilitarian social welfare function that is indifferent to inequality they assume that individuals have diminishing marginal utility of income, which restores the value of making incomes more equal.

Kotlikoff. Kotlikoff calls his proposal the "Fair Tax," one that would replace all the federal taxes, including the payroll tax for Social Security. It consists of a 30% retail sales tax rate, equivalent to a 23% tax on income, along with a rebate to protect low-income taxpayers that is a function of income and personal characteristics.[2] The Fair Tax also requires some cuts in expenditures to maintain budget neutrality.

- A combination of a cash-flow tax on business and a wage tax on individuals. The cash flow of a business equals the firm's revenues less all its expenditures, consisting of material inputs, investments, and wages and salaries. Think of it as the European-style consumption value added tax, which taxes revenues less expenses for material inputs and investments, from which wages and salaries are also deducted from the tax base. The wage component is then picked up at the personal level, where it can be levied on a progressive basis. Since investment is excluded from tax, and investment equals saving, the combined tax is a consumption tax. Robert Hall and Alvin Rabushka were the first to propose a consumption tax of this form. They favor a flat tax of 19% on wages less a personal exemption that varies with family characteristics (Hall and Rabushka, 1995). David Bradford proposed a variation that he called the X tax, which levies graduated tax rates on the personal wage tax to achieve progressivity, and taxes business cash flow at the highest marginal tax rate applied to wages to discourage taxpayers from shifting income between business and personal accounts to avoid taxes (Bradford, 1986).

The Kotlikoff, Hall/Rabushka, and Bradford proposals have one decided advantage for individuals over the first two proposals, that of greatly simplifying the payment of taxes. The Kotlikoff retail sales tax essentially removes the taxpayer from any interaction with the IRS, except for the processing of the rebates. And the tax should be easy to implement since 45 states already levy their own retail sales taxes; the federal government can exploit the collection mechanisms already in place. The Hall/Rabushka and Bradford proposals do not require individuals to keep track of their saving and asset holdings, and a wage tax can be filed on a postcard consisting of just a few lines. Taxpayers report their wage and salary income on one line, subtract any rebates or exemptions, compute the tax liability on their net income from a tax table, and compare their tax liability with the taxes withheld on their wage and salary income to determine whether they owe more taxes or should receive a refund.

2. The 30%/23% equivalence comes from the formula $(1 + t)(1 + t^*) = 1$ for equivalent broad-based taxes developed in Chapter 16. $(1 + 0.3)(1 - 0.23) = (1.3)(0.77) = 1$. The Fair Tax is described and analyzed in Jokisch and Kotlikoff, 2005.

Beyond the question of simplicity, what does tax theory say about these various broad-based takes that might help choose among them? At the most basic level, the answer is not enough to make a choice. Previous chapters developed two theoretical results relating to broad-based taxes, neither of which is entirely helpful.

The first result is the optimal commodity tax formula developed in Chapter 13, reproduced here as Eqn (17A.1).

$$\frac{\sum_{i=1}^{N} t_i M_{ki}}{M_k} = C, \quad k = 2, \ldots, N \tag{17A.1}$$

Recall that the commodities can be both goods and factors. The formula was developed in a static framework but is valid as well over time, in which the relevant substitution effects across goods and factors would apply both within any given year and over time. We have very little knowledge of most of these compensated cross-price elasticities, but the implication of the formula is that broad-based taxation of any kind is not optimal, in general. This is underscored by the Corlett–Hague analysis, also in Chapter 13, which says that efficiency loss from a broad-based proportional tax can be reduced by a marginal, revenue-neutral change that increases (decreases) tax rates on goods relatively more (less) complementary to leisure. Hence, the comparison of broad-based taxes is a comparison of decidedly third-best options, with no good knowledge of how far away from optimal any broad-based proposal might be.

The second result is the equivalence of broad-based taxes developed in Chapter 16: In a static, perfectly competitive, profitless economy with identical individuals, all broad-based taxes are equivalent in terms of their incidence. There would be nothing to choose between income and consumption taxes, or whether individuals or businesses were taxed. One might add an implicit assumption that went unstated in the chapter, that the government does not change its behavior in response to different taxes so that there is no effect on the resulting general equilibrium from the expenditure side of the budget. As pointed out in Chapter 16, the differences between various broad-based taxes are in the details, and the details matter. A short list relevant to the tax reform literature would include the introduction of time in thinking about lifetime incidence; the possibility of nonlinear taxes, either through graduated tax rates or a single tax rate (a so-called flat tax) combined with a rebate or exemption (both the optimal tax result and the broad-based equivalence results assume linear taxes); the fact that individuals vary along a number of dimensions such as tastes, skills, and their position in the life cycle, particularly whether they are working or retired; future incomes are uncertain; and the existence of various market imperfections such as market power and credit constraints.

All these factors can influence the choice among broad-based taxes. Nonetheless, the equivalence of broad-based taxes in the simplest baseline model is worth keeping in mind. One recalls Richard Musgrave's belief, noted in Chapter 11, that the choice between income and consumption as the tax base is far less important than how progressive either tax would be. In his view, the amount of vertical equity or distributive justice inherent in the tax structure is more important than the efficiency implications of choosing between income and consumption as the tax base.

Musgrave's position notwithstanding, efficiency implications have captured the majority of economists' attention in assessing proposals such as these, with efficiency defined over the long-run steady state following a tax reform. The most common comparisons among reform proposals are done with perfect foresight, perfectly competitive, full-employment, OLG models in which individuals are life-cycle consumers, models pioneered by Alan Auerbach and Kotlikoff. We discussed this approach in Chapter 17 and will only recall the main results here:

1. The personal consumption tax generates the largest steady-state gains in output and individual welfare.
2. The transition to the steady state matters. A large portion of the gains in moving from an income to a consumption tax arise because a consumption tax taxes the current wealth holdings of the middle aged and elderly at the time of the change, making this component of the tax a lump-sum tax. Consequently, both groups lose during the transition, which is a primary reason the United States has not adopted a personal consumption tax or the other variations of consumption taxation mentioned above. By transferring resources initially from the relatively high-MPC middle aged and elderly to the relatively low-MPC young, saving and investment increase, which increases consumption per person available to everyone in the long run. Note that this outcome depends importantly on the state of the economy at the time of the reform. In an OLG framework, the US government can redistribute across generations to ensure that the level of saving and investment is consistent with the Golden Rule of Accumulation, in which consumption per person is maximized. The United States has not done this; the capital stock per person is below the Golden Rule level. Therefore, consumption taxation generates the redistribution from older to younger generations that brings the economy closer to the Golden Rule stock of capital. Were the economy there at the time of the reform, the movement to consumption taxation would not be as beneficial.
3. Attempts to protect the middle aged and elderly during the transition by, say, allowing them to continue to depreciate their existing assets remove the majority of the gains from adopting a personal consumption tax. Similarly, attempts to protect the poor within the tax system, such as by the rebates or exemptions in the Kotlikoff and Hall/Rabushka proposals, also remove much of the gains because they require higher tax rates. In summary, even these kinds of OLG analyses do not establish a clear winner among the four main reform proposals.[3]

Should Income from Capital be Taxed?

We turn now to a number of issues relating to tax reform that have not been covered previously.[4] The first is the large literature on whether income from capital should be taxed.

The argument against taxing income from capital most often rests on two results. One is due to Anthony Atkinson and Joseph Stiglitz, referenced in Chapter 15 (Atkinson and Stiglitz, 1976). They consider nonlinear taxation in a static model in which individuals have utility defined over a number of consumer goods and labor. They show that if labor is weakly separable from all the consumer goods, then the consumer goods should be taxed at the same rate. Equivalently, only labor income need be taxed. The intuition is that, under weak separability, the government cannot achieve any distributional goals with differential taxation of the consumer goods that it cannot achieve with a tax on labor. If we think of the consumer goods as a single good purchased in different time periods, with labor earned in the first period, then there should be no tax on the rate of interest that links the goods through time. That is, there should be no tax on income from capital.

The second result is due separately to Kenneth Judd and Christophe Chamley. They show that, in a representative consumer, infinite horizon Ramsey model with linear taxes, the tax on income from capital should be zero in the long run. The intuition is that if the rate of interest equals the marginal product of capital, then the MRT from one period to the next equals $(1 + r)$, whereas with taxation of interest, the MRS is $1 + r(1 - t)$. Since interest income is taxed each

3. An excellent study of this kind is Altig et al., 2001. It compares variations of all the main reform proposals except for Kotlikoff's Fair Tax.

4. Readers interested in the issues discussed from here to the end of the appendix should definitely consult Boadway, 2012. It is a superb, comprehensive treatment of the relationship between tax theory and tax reform. A second highly informative source on issues related to the choice of a tax base is J. Banks and P. Diamond, 2010 "The Base for Direct Taxation," which they wrote as a background study for the Mirrlees Review of the British tax system. It is published in Mirrlees, James, Stuart Adam, Timothy Besley, Richard Blundell, Stephen Bond, Robert Chote, Malcolm Gammie, Paul Johnson, Gareth Myles, and James Poterba (eds), *Dimensions of Tax Design*, 548–648. Oxford University Press, Oxford, UK, 2010. The question of whether income from capital should be taxed is addressed at length by Banks and Diamond.

period, at time T the wedge between the MRT and MRS grows to $\left[\frac{1+r}{1+r(1-t)}\right]^T$, which becomes large without limit as $T \to \infty$.

One might dismiss the Atkinson/Stiglitz result on the grounds that labor is almost certainly not weakly separable, but the Judd/Chamley result is more difficult to ignore. It is not clear why the relationship between consumption today and in the distant future should be so heavily distorted (Chamley, 1986; Judd, 1985).

There is also the argument noted above that is commonly used to justify consumption rather than income taxes, that not taxing income from capital promotes saving and investment and moves the US economy closer to the consumption-maximizing Golden Rule of Accumulation in the steady state.

More recently, a number of articles have appeared that argue in favor of taxing income from capital, enough so that the tide seems to be shifting in that direction. The arguments relate to the details mentioned above that tend to upset the equivalence of broad-based taxes. We will note two of them by way of illustration.

If the Atkinson/Stiglitz framework is extended to include the earning of income in multiple periods, then future incomes may well be uncertain. In that case, taxing any component of income can provide an insurance function under progressive taxation that has value to individuals. Taxes rise proportionally more than income in periods with particularly good income draws, and fall proportionately more than income in periods with particularly bad draws, thereby smoothing income and consumption over time. Individuals prefer to smooth consumption over time if they have diminishing marginal utility of income. (We will pursue the desire for consumption smoothing more closely in Chapter 20.)

A second argument is due to Emmanuel Saez. He has found that the propensity to save is directly related to an individual's ability or skills: higher skilled individuals want to save a higher proportion of their incomes than do lower skilled individuals. Equivalently, higher skilled individuals act as if they discount the future at a lower rate. If this is true, then taxing income from capital is a way of increasing redistribution through taxation, which is desirable if society wants a progressive tax system (Saez, 2002).

Arguments such as these in favor of taxing income from capital are still mindful of the Judd/Chamley result that this implies increasing distortion of present and future consumption over time. Consequently, economists no longer believe that income from labor and capital should necessarily be taxed at the same rate. The modern view rejects the Haig−Simons notion that all income represents an increase in purchasing power and should therefore be taxed at the same rate. The increasing inefficiency over time associated with taxing income from capital argues in favor of taxing income from capital at a lower rate than income from labor. No consensus has formed as yet on how different the rates should be.[5]

Here is a final practical point to consider. Suppose, after considering all the arguments, society decides to reform its current income tax to remove the taxation of capital but retain the tax on individuals. There are a number of ways to do this and they have different implications. One way is to exempt all saving from taxation, exempt any returns on the assets as the returns accrue, and then tax assets as they are withdrawn. This is referred to as the EET method: Exempt the saving, Exempt the returns as they accrue, and Tax the withdrawals. The EET method turns the income tax into a personal consumption tax, with saving exempt from taxation. The IRAs in the United States are taxed on this basis to encourage people to save for their retirement. Another way is to tax (not to exempt) the savings, but then exempt any returns to assets as they accrue and exempt the withdrawals from the assets. This is referred to as the TEE method: Tax the saving, Exempt the returns as they accrue, and Exempt the withdrawals. Since the returns to capital and the withdrawals are exempt, this turns the income tax into a wage tax (ignoring income from land rents, which might also be taxed). The income from capital is never taxed. The Roth IRAs in the United States are taxed on this basis.

The EET and TEE methods would be equivalent in a world of perfect certainty in which all returns to capital are entirely predictable in advance. Under these conditions, the value of an investment would just equal the anticipated present value of its returns over time. Therefore, taxing that value initially under the TEE or at withdrawal under the EET is a matter of indifference to individuals. The two methods differ with uncertain returns, however. The EET captures unanticipated abnormal returns (or losses), whereas the TEE method does not.

The EET and TEE methods also have quite different effects in the Auerbach−Kotlikoff OLG framework. The EET method transfers resources from the older to the younger generations because of the taxation of existing wealth as the elderly draw down their wealth to consume in retirement. As noted in Chapter 17, Auerbach and Kotlikoff refer to this as an investment incentive since it favors new capital over old capital. The TEE method transfers resources from the younger to the older generations since income from existing capital is untaxed after the reform. The TEE method does increase saving and investment over time, but it is not as effective as the EET method at increasing welfare over the long run because it does not enjoy the advantage of the lump-sum tax on the wealth of

5. The Nordic countries have generally adopted a dual rate income tax, with capital income taxed at a lower rate than labor income.

the middle-aged and older generations alive at the time of the reform.

The British recently engaged in a review of their entire tax system under the direction of James Mirrlees. The Mirrlees Review recommended a variation of these two methods that is referred to as TtE: Tax the savings, allow a deduction as the returns accrue equal to the safe return on the assets as measured by a predetermined short-term interest rate (as opposed to deducting/exempting all the accruing returns), and then Exempt the withdrawals from the assets. This is a tax on wage income plus any abnormal returns to capital (less any abnormal losses). In fact, the Mirrlees Review recommended the use of all three methods: EET for pension savings accounts since they were already taxed on that basis, TEE for bank accounts to capture the otherwise untaxed services these accounts offer, and TtE for all other assets whose value is above a certain threshold amount (those below the threshold can use the TEE method).[6]

Classical versus Newer Tax Theory

In *From Optimal Tax Theory to Tax Policy: Retrospective and Prospective Views*, Robin Boadway makes a useful distinction between the older "classical" tax theory and the newer tax theory.[7] The classical theory operates under the assumption of perfect information and assumes the government has a fixed number of tax instruments at its disposal. The issue is how to use these given instruments to maximize social welfare. The derivation of optimal commodity taxes, Eqn (17A.1), is in the classical tradition.

The newer theory begins with the assumption of imperfect information, which raises two new sets of issues. The first is the introduction of incentive compatibility constraints into the government's maximization problem, so that high-ability people will not pretend to be low-ability people to avoid paying taxes. These additional constraints are necessary because the government cannot directly observe individuals' abilities or skill levels, only their incomes. We developed this point in Chapter 15.

Imperfect information gives a new interpretation to results such as the Corlett–Hague analysis and the Saez finding that saving is related to ability: they serve to relax the incentive compatibility constraints and permit more redistribution to occur. Regarding the Corlett–Hague analysis, by raising taxes on goods complementary to leisure, high-ability individuals have less incentive to pretend to have low ability and take more leisure time, since leisure activities are now more expensive. Similarly, if the propensity to save is directly related to ability, high-ability individuals who pretend to have low ability want to save more than individuals who really do have low ability. With saving (income from capital) now taxed, this raises the costs of pretending to have low ability.

A second important implication of imperfect information that is not highlighted in the text is that it leads to a search for new things to tax. The tax instruments are no longer fixed as in the classical theory. Instead, tax theory has a mechanism design aspect to it in line with modern contract theory under imperfect information. If a contract cannot precisely specify the item that is the subject of the contract because of poor information, then in general it pays to add features to the contract that are related to the item of interest even if the relationship is only imperfect. The same principle applies to taxation. Think of the original Mirrlees optimal income tax model. The item of interest is the skill level of the different individuals. If the skills were known with certainty, then taxes could be based on skills. They would be lump sum and, given the standard social welfare function, would imply leveling everyone to the mean income. Since skills are unknown, the government has to tax income, which generates inefficiencies and reduces the amount of redistribution that is possible. More redistribution can be achieved by taxing attributes related to skills or income, which is certainly related to skills. One example is Saez's call to tax saving.

Varying Tax Rates by Age

Michael Kremer, in a highly influential paper, has argued that tax rates should vary with age, both to increase the redistribution possible through the income tax and to reduce its inefficiency. His argument rests on three points.

Labor supply elasticities by age—The evidence is not airtight, but Kremer believes that young people, which he defines as those aged 17–21 years, have much higher labor supply elasticities than older workers, which he defines as those aged 31–64 years. Two pieces of supporting evidence are that the young move in and out of employment/unemployment more than the middle aged and that the middle aged are more likely to have salaried jobs for which the hours of work cannot be varied. Consider raising the marginal tax rate on income level x within each of the two age groups. The dead-weight loss from raising the tax rate will be much higher for the younger than the older workers. Therefore, there is an efficiency gain from having a lower marginal tax rate at income x for the younger workers and raising it in a revenue-neutral manner for the older workers.

The income distribution by age—The distribution of income becomes much more unequal as cohorts age.

6. For these and other issues related to taxing income or consumption, see Auerbach, 2006. See, also, Auerbach's analysis of the Mirrlees Review recommendations in Auerbach, 2012, pp. 685–708.
7. R. Boadway, *op. cit.*, Chapter 2.

Consider the so-called hazard rate at income x for each age group, defined as $\frac{f(x)}{1-F(x)}$, where F is the cumulative density function of income. The hazard rate is the proportion of individuals at income x divided by the proportion of individuals with incomes greater than x. According to Kremer, the hazard rate is over five times greater for the 17- to 21-year age group than for the 31- to 64-year age group over a wide range of incomes because the older age group has so many more individuals who are earning incomes greater than the chosen income. And it is over two times larger for the 21- to 26-year age group. This matters, because raising the marginal tax rate at income x is an inframarginal event for all those with incomes greater than x. It raises tax revenue without any efficiency loss because the response to the higher tax rate depends only on the income effect. Since the income effect tends to be positive for labor supply, this implies even more labor supply and tax revenue collected from those with higher income than if there were no income effect. This revenue raising effect is yet another argument for raising the marginal tax rate on x for the older workers.

The correlation of incomes over age groups—Income earned at the younger ages is almost completely uncorrelated with income earned at the older ages. Consequently, raising marginal tax rates on older workers and lowering them on younger workers redistributes lifetime incomes, which it would not do if incomes across age groups were perfectly correlated. This is the third argument for varying tax rates by age.[8]

The principle that taxes should be related to attributes that vary with skills or income has to be tempered by what people will find acceptable. Height is also strongly positively correlated with income, but people might well balk at varying marginal tax rates by height even if they accept tax rates that vary by age. Taxing people by height singles out particular people to pay higher or lower taxes, whereas varying taxes by age applies equally to everyone over time. Also, the Social Security pension system does incorporate age in a number of ways in determining pensions, so people in the United States have some familiarity with age varying taxes and transfers (see Chapter 20).[9]

Commitment

The ability of a government to commit to its tax policies over time is a central issue in a world of imperfect information. To give a common example, governments always have an incentive to tax existing wealth because it represents a sunk cost and thus the tax is lump sum. Suppose the government announces a tax on existing wealth and promises that it will be a one-time tax. The government has a strong incentive to renege on its promise. Once individuals invest in new assets this period, the wealth embodied in those assets becomes a sunk cost next period and another source of lump-sum tax revenue for the government. The government's one-time wealth tax policy is said to be time inconsistent.

A more natural outcome is one in which the government and the taxpayers play a time-consistent principal-agent game, with the government as the principal and the taxpayers as the agents. The government announces its tax policy first and then the taxpayers react to the policy. The taxpayers, knowing that the government cannot commit to a one-time wealth tax, will reduce their investment in anticipation of paying higher taxes on those assets in the second period. The government understands this, and designs its policies over time knowing how the taxpayers will respond to its policies. The resulting equilibrium is time consistent, but it involves much higher taxes and much lower investment than would exist if the government could plausibly commit to a one-time wealth tax.[10]

The problem of commitment generalizes to all second-best tax policies set in a world of imperfect information. Suppose the government designs a tax policy with incentive compatibility constraints that work as intended, with all individuals correctly self-selecting according to their abilities or skills. Once the tax is in place and individuals self-select, the government knows who the higher and lower ability people are. It then has an incentive to renege on its original policy and levy taxes thereafter on the basis of the revealed abilities, a lump-sum tax. Most theoretical analysis of tax policy under imperfect information, such as the Stiglitz analysis presented in Chapter 15, simply assumes that the government can commit to its policies, an uncomfortable position to adopt. The only saving grace is that governments typically do commit to their tax policies for long periods of time, for reasons that are not well understood.

REFERENCES

Aaron, H., McGuire, M., November 1970. Public goods and income distribution. Econometrica 38 (6), 907–920.

Altig, D., Auerbach, A., Kotlikoff, L., Smetters, K., Walliser, J., June 2001. Simulating fundamental tax reform in the United States. American Economic Review 91 (3), 574–595.

Aronson, J., Lambert, P., June 1994. Decomposing the gini coefficient to reveal the vertical, horizontal, and reranking effects of income taxation. National Tax Journal 47 (2), 273–294.

8. Kremer, 2001. The hazard rate data are on p. 1.

9. Unfortunately, Social Security has it backward, levying higher effective marginal tax rates on the young and some of the elderly. We will pursue this point in Chapter 20.

10. Switching from an income to a consumption tax is one way to tax existing wealth that the government is more likely to commit to.

Aronson, J., Johnson, P., Lambert, P., March 1994. Redistributive effect and unequal income tax treatment. Economic Journal 104 (423), 262–270.

Atkinson, A., 1994. The distribution of the tax burden. In: Quigley, J., Smolensky, E. (Eds.), Modern Public Finance. Harvard University Press, Cambridge, MA chapter 2.

Auerbach, A., Kotlikoff, L., 1987. Dynamic Fiscal Policy. Cambridge University Press, New York.

Altig, D., Auerbach, A., Kotlikoff, L., Smetters, K., Walliser, J., June 2001. Simulating fundamental tax reform in the United States. American Economic Review 91 (3), 574–595.

Atkinson, A., Stiglitz, J., July–August, 1976. The design of tax structure: direct versus indirect taxation. Journal of Public Economics 6 (1–2), 55–75.

Auerbach, A., June 2006. The Choice between Income and Consumption Taxes: A Primer. NBER Working Paper, No. 12307.

Auerbach, A., September 2012. The mirrlees review: a U.S. perspective. National Tax Journal 65 (3), 685–708.

Barnett, J., Vidal, P., July 2013. State and Local Government Finances Summary: 2011, Appendix Table A.1. U.S. Bureau of the Census website. www2.census.gov/govs/local/summary.pdf.

Barro, R., November/December, 1974. Are government bonds net wealth? Journal of Political Economy 82 (6), 1095–1117.

Blinder, A., 1980. The level and distribution of economic well-being. In: Feldstein, M. (Ed.), The American Economy in Transition. University of Chicago Press, Chicago.

Browning, E., Johnson, W., 1979. The Distribution of the Tax Burden. American Enterprise Institute for Public Policy Research, Washington, DC.

Supplement Budget of the United States Government, Fiscal Year 2014, 2014. U.S. Government Printing Office, Washington, DC. part Five: Historical Tables, Table 2.1.

Banks, J., Diamond, P., April 2010. The Base for Direct Taxation, Prepared for the Report of a Commission on Reforming the Tax System for the 21st Century, Chaired by James Mirrlees, The Institute for Fiscal Studies. In: James, M., Adam, S., Besley, T., Blundell, R., Bond, S., Chote, R., Gammie, M., Johnson, P., Myles, G., Poterba, J. (Eds.), Dimensions of Tax Design. Oxford University Press, Oxford, UK, pp. 548–648.

Boadway, R., 2012. From Optimal Tax Theory to Tax Policy: Retrospective and Prospective Views. MIT Press, Cambridge, MA.

Bradford, D., 1986. Untangling the Income Tax. Harvard University Press, Cambridge.

Chirinko, R., December 1993. Business fixed investment spending. Journal of Economic Literature 31 (4), 1875–1911.

Chirinko, R., Fazzari, S., Meyer, A., October 1999. How responsive is business capital formation to its user cost? an exploration with micro data. Journal of Public Economics 74 (1), 53–80.

Chamley, C., May 1986. Optimal taxation of capital income in general equilibrium with infinite lives. Econometrica 54 (3), 607–622.

Davies, J., Whalley, J., March 1989. Taxes and Capital Formation: How Important Is Human Capital?. NBER Working Paper #2899 National Bureau of Economic Research, Cambridge, MA.

Davies, J., St Hilaire, F., Whalley, J., September 1984. Some calculations of lifetime tax incidence. American Economic Review 74 (4), 633–649.

Diamond, P., December 1965. National debt in a neoclassical growth model. American Economic Review 65 (5:1), 1126–1150.

Ebert, U., Moyes, P., January 2000. Consistent income tax structures when households are heterogeneous. Journal of Economic Theory 90 (1), 116–150.

Engen, E., Gale, W., May 1997. Consumption taxes and saving: the role of uncertainty in tax reform. American Economic Review 87 (2), 114–119.

Fullerton, D., Rogers, D., 1993. Who Bears the Lifetime Tax Burden? The Brookings Institution, Washington, DC.

Gillespie, W., 1965. Effect of public expenditures on the distribution of income. In: Musgrave, R. (Ed.), Essays in Fiscal Federalism. The Brookings Institution, Washington, DC.

Goolsbee, A., February 1998. Investment tax incentives, prices, and the supply of capital goods. Quarterly Journal of Economics 113 (1), 121–148.

Hines, J., March 2000. What is benefit taxation? Journal of Public Economics 75 (3), 483–492.

Hall, R., Rabushka, A., 1995. The Flat Tax, second ed. Hoover Institution Press, Stanford, CA.

Jokisch, S., Kotlikoff, L., December 2005. Simulating the Dynamic Macroeconomic and Microeconomic Effects of the Fair Tax. NBER Working Paper No. 11858.

Judd, K., October 1985. Redistributive taxation in a simple perfect foresight model. Journal of Public Economics 28 (1), 59–83.

Kakwani, N., March 1977. Measurement of tax progressivity: an international comparison. Economic Journal 87 (345), 71–80.

Kaplow, L., June 1989. Horizontal equity: measures in search of a principle. National Tax Journal 42 (2), 139–154.

Keen, M., Papapanagos, H., Shorrocks, A., January 2000. Tax reform and progressivity. Economic Journal 110 (460), 50–68.

Kiefer, D., December 1984. Distributional tax progressivity indexes. National Tax Journal 37 (4), 497–513.

Kiefer, D., January 1991. A comparative analysis of tax progressivity in the United States: a reexamination," plus "comments" and "reply," Public Finance Review 19 (1), 94–108.

Kotlikoff, L., December 1984. Taxation and savings: a neoclassical perspective. Journal of Economic Literature 22 (4), 1576–1629.

Kotlikoff, L., Summers, L., 1985. Tax incidence. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, vol. II. North-Holland Elsevier, New York, chapter 17.

Kremer, M., March 6, 2001. Should Taxes Be Independent of Age?. Working Paper Littauer Center, Harvard University, Cambridge, MA.

Lambert, P., September 1993a. Evaluating impact effects of tax reforms. Journal of Economic Surveys 7 (3), 205–242.

Lambert, P., 1993b. The Distribution and Redistribution on Income: A Mathematical Analysis, second ed. University Press, Manchester.

Maital, S., 1975. Apportionment of public goods benefits to individuals. Public Finance 30 (3), 397–416.

McClure, C., Thirsk, W., June 1975. A simplified exposition of the harberger model, II: expenditure incidence. National Tax Journal 28 (2), 195–207.

Metcalf, G., December 1999. A distributional analysis of green tax reforms. National Tax Journal 52 (4), 655–681.

Moyes, P., Shorrocks, A., July 1998. The impossibility of a progressive tax structure. Journal of Public Economics 69 (1), 49–65.

Nerlove, M., Razin, A., Sadka, E., von Weizsacker, R., March 1993. Comprehensive income taxation, investments in human and physical capital, and productivity. Journal of Public Economics 50 (3), 397–406.

Pechman, J., 1985. Who Paid the Taxes, 1966–85? The Brookings Institution, Washington, DC.

Pechman, J., Okner, B., 1974. Who Bears the Tax Burden? The Brookings Institution, Washington, DC.

Scarf, H., 1967. On the computation of equilibrium prices. In: Ten Economic Studies in the Tradition of Irving Fisher. Wiley, New York.

Scarf, H., September 1969. An example of an algorithm for calculating general equilibrium prices. American Economic Review 59 (4:1), 669–678.

Shoven, J., Whalley, J., September 1984. Applied general equilibrium models of taxation and international trade. Journal of Economic Literature 22 (3), 1007–1051.

Saez, E., February 2002. The desirability of commodity taxation under non-linear income taxation and heterogeneous tastes. Journal of Public Economics 83 (2), 217–230.

Trostel, P., April 1993. The effect of taxation on human capital. Journal of Political Economy 101 (2), 327–350.

Whalley, J., November 1984. Regression or progression: the taxing question of incidence. Canadian Journal of Economics 17 (4), 654–682.

Chapter 18

The Second-Best Theory of Public Expenditures: Overview

Chapter Outline

The second-best public expenditure theory has been at the forefront of theoretical developments in public sector economics for the past 50 years. The latest wave of research is exploring the effects of private information on public sector decision rules, just as private information has dominated recent developments in tax theory.

In extending the methodology of second-best tax theory to expenditure theory, public sector theorists have shattered the received doctrine of first-best expenditure theory that is still featured in the undergraduate textbooks. No longer is it possible to accept, even as approximations, such time-honored decisions rules as $\Sigma MRS = MRT$ for externalities, and marginal cost pricing for decreasing cost services, rules that bear close intuitive relationships to competitive markets. As it now stands, public expenditure theory is more than a little chaotic, each new journal article pushing in new directions and offering new insights, with little in the way of synthesis to provide some clues as to where the second-best expenditure theory will eventually lead. Perhaps this is as it must be.

The essence of second-best analysis is the addition of constraints to the basic first-best general equilibrium model beyond the fundamental constraints of production technologies and market clearance (and possibly resource limitations). The promise of the second-best theory is its move toward realism that the additional constraints capture important features of actual economies such as the existence of monopoly elements in the private sector, distorting taxes, and private information. The drawback of the theory is that the second-best decision rules necessarily vary depending upon both the number of constraints added to the model and the precise form that each constraint takes. Unlike the first-best analysis, then, the set of policy prescriptions is virtually unlimited. Furthermore, no second-best theory can possibly incorporate all the additional constraints that would be necessary to approximate reality. As was noted in Chapter 12, current second-best models accurately portray only a very few specific distortions operating in the economy. They model the remaining parts of the economic system along standard first-best lines. Hence, the state of the art is still but a hesitant first step or two toward reality despite 50 years of analysis. Even so, virtually none of the old first-best decision rules has remained standing. Needless to say, the goal of developing a widely accepted normative economic theory of the public sector appears increasingly less plausible.

A few chapters in a broad text such as this one cannot do justice to all the ways that public sector theorists have chosen to rework public expenditure theory in a second-best context, much less provide a comprehensive synthesis of this large and varied literature. In lieu of this, the next few chapters undertake a far more modest task. We merely want to highlight a few of the principal second-best public expenditure results to date and to demonstrate the more common methodological tools used to analyze public expenditure theory in a second-best context.

The topics have been chosen to lend some coherence to the overall text. We hope to achieve this by limiting the analysis for the most part to the specific public expenditure problems discussed in Part II under first-best theory. In addition, we will concentrate on the two most common second-best constraints employed in the literature. One is the existence of private information. The other is the government's need to rely on distorting unit "commodity" taxes (subsidies) on consumer goods and factors to finance its expenditures. The market environment, in contrast, is assumed to be perfectly competitive and therefore first best, unless specifically stated otherwise. We will also usually

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00018-2

assume that the government's budget must balance. These additional assumptions are commonly employed in the second-best literature.

Structuring the public expenditure models in this way allows us to draw directly upon the models developed in Chapters 13−16 for analyzing the second-best theory of taxation. More often than not, these tax models require only slight modifications to incorporate public expenditure questions. This is why it made sense to reverse the development of Part II and consider second-best tax theory before second-best expenditure theory. It is natural to introduce distorting taxes as the additional constraint necessitating a second-best approach to public expenditure questions. The tax models also prove convenient as an analytical framework for developing two of the strongest results in all of second-best theory. The first is that the optimal commodity tax rules are unaltered by the presence of government expenditures. The second is that the second-best public expenditure decision rules tend to have their most appealing and simplest interpretations when distorting taxes are set optimally to maximize social welfare or minimize loss.

Regarding methodology, some of the expenditure problems will be analyzed from the perspective of social welfare maximization, while others from the perspective of utility maximization or loss minimization of a representative consumer. The latter is especially useful when distributional considerations are not central to the point being developed. Switching the analytical framework in this manner will allow us to demonstrate a variety of models suitable for analyzing second-best expenditure questions. It should not be confusing since both approaches have been fully explored in the preceding chapters on second-best tax theory. With these goals in mind, a selection of second-best public expenditure results will be presented in six self-contained chapters.

Chapter 19 analyzes transfer payments under second-best assumptions, with an emphasis on the ways in which private information affects the optimal design of government transfer programs. We saw in Chapter 17 that the theory of distorting transfers under perfect information is essentially subsumed within second-best tax theory, because transfers are analytically equivalent to negative taxes. Not so the theory of transfers under private information. Private information raises many important issues with the design of transfer programs that are absent with taxation. Taxpayers are particularly concerned that transfer recipients might use their private information to undermine the intent of the transfer programs, such as able-bodied people accepting public assistance instead of working. Much of the recent theoretical work on transfers has focused on the mechanism design problem of preventing potential recipients from misrepresenting themselves to the government. Two central design issues are whether transfers should be cash or in kind to prevent cheating and whether transfers can be decentralized or must instead be provided directly by the government. The chapter also considers some second-best design issues associated with transfers that are unrelated to private information.

Chapters 20 and 21 cover entirely new material, the provision of public or social insurance. Chapter 20 considers medical insurance and Chapter 21 considers Social Security pensions. We did not cover public insurance previously because private information is the main reason why some forms of insurance are driven into the public sector. That is, social insurance is inherently a second-best topic. In fact, all governments in the highly industrialized market economies are heavily involved in the provision of insurance, particularly medical insurance and public pensions. We will see that simply having the government provide insurance does not eliminate many of the problems that make private insurers reluctant to provide the insurance.

Chapter 22 reworks the first-best theory of externalities contained in Chapters 6−8. The chapter begins with the question of how distorting taxation affects standard first-best decision rules, under the assumption of perfect information. In a first-best environment, whether one considers a nonexclusive Samuelsonian public good or exclusive activities that generate either "individualized" or "aggregate" externalities, we saw that the government should achieve an allocation in which $\sum_{h=1}^{H} \text{MRS}^h = \text{MRT}$.[1] For exclusive goods, the government can, in principle, tax (subsidize) the externality-generating activity to achieve the desired result.

The summation rule fails to hold for any of these cases in a second-best environment. Chapter 22 will demonstrate this for Samuelsonian nonexclusive goods when the revenues to pay for this good must be raised with distorting commodity taxes. This problem was first considered by A. C. Pigou in 1947 but formalized more precisely in the 1970s, first by Anthony Atkinson and Nicholas Stern, and then separately by Peter Diamond and David Wildasin (Pigou, 1947; Atkinson and Stern, 1974; Diamond, 1975; Wildasin, 1984). The main question of interest is whether applying the first-best $\Sigma\text{MRS} = \text{MRT}$ rule would lead to too much or too little of the public good when taxes are distorting.[2]

1. This applies, of course, only to consumption externalities. The standard production-externality rule is that $\text{MRS} = \sum_{j=1}^{J} \text{MRS}^j$ for J firms.

2. Leuthold derived the second-best tax rules for an aggregate externality when society cannot redistribute lump-sum to satisfy the first-best interpersonal equity conditions. He showed that the failure to satisfy interpersonal equity destroys the optimal properties of competitive markets even for purely private goods that are not generating any external effects. The government must tax (subsidize) these goods as well. The pervasiveness of government intervention in the market economy is obviously a discouraging result for capitalist societies. See Leuthold (1976). Also see Hartwick (1978).

The effect of private information on externalities is difficult to characterize because it depends upon the nature of the private information and the form of the externality. Rather than attempt a comprehensive analysis, the chapter shows how private information can alter the efficiency properties of two time-honored externality policy prescriptions for externalities from first-best theory, Pigovian taxes and Coasian bargains.

Chapter 23 turns to a well-known second-best result for decreasing-cost industries. We saw in Chapter 9 that these services should be provided at marginal cost prices in a first-best policy environment, with lump-sum taxes covering the resulting losses to the firm. We noted, however, that these services are typically priced at average cost in the United States, and considered a number of equity issues arising from this practice. Chapter 21 expands upon the average-cost pricing philosophy by applying it to a multiproduct firm. Marcel Boiteux wrote the classic article on this issue in 1956. He considered the pricing and investment implications of requiring that a multiservice decreasing-cost industry cover its full costs out of total revenues. Price does not necessarily equal average cost for each service,[3] but the prices on all the services combined must raise enough total revenue to cover all costs. The Boiteux problem is especially intriguing for public sector economics because it closely parallels the optimal commodity tax problem of Chapters 13 and 14 in which the government has to set taxes (consumer prices) to collect a given amount of revenue. As we shall see, the optimal pricing rules for the multiproduct decreasing-cost firm have virtually the same interpretation as the optimal tax rules. They also apply to any government agency that operates under a legislated budget constraint, whether or not the agency's output exhibits decreasing cost production.

Chapter 24 concludes the second-best public expenditure analysis by considering the general problem of government production in a second-best environment. There are no specific constraints on government production possibilities; they can exhibit decreasing, increasing, or constant returns to scale. Furthermore, the government is permitted to produce anything that the private sector produces. The analysis places only two realistic constraints on government activity. First, if government producers buy and sell inputs and outputs, they must do so at the (competitive) prices faced by the private sector producers. Otherwise, they may not be able to compete effectively with the private sector for scarce inputs or the sale of their outputs. Second, if government production incurs a deficit (surplus) at these prices, the government must use distorting commodity taxes (subsidies) to cover the deficit (return the surplus). This type of general expenditure model was first explored in depth in the 1970s, most notably by Peter Diamond and James Mirrlees, and Robin Boadway (Diamond and Mirrlees, 1971; Boadway, 1975, 1976). Diamond and Mirrlees asked the following questions in their seminal article entitled "Optimal Taxation and Public Production":

1. Does the existence of government production alter the optimal commodity tax rules derived in the context of raising revenue simply for the sake of raising revenue? The answer turned out to be no.
2. What production rules should the government follow in the presence of optimal commodity taxation to cover production deficits (return surpluses)? Their answer to this question was especially surprising. They showed that if the distorting taxes are optimal, then the government should follow standard *first-best* production rules.

Boadway generalized the analysis to consider the welfare effects of raising additional taxes (subsidies) and/or marginally increasing government production from *any* initial values of distorting taxes and government production, not necessarily the optimal values. As might be imagined, the resulting tax and expenditure rules are extremely complex. One nice result, however, is that the addition of government production does not affect the nonoptimal marginal tax loss rules developed in Chapters 13 and 14.

Chapter 25 concludes Part III with a discussion of the various anomalies uncovered in the new and rapidly growing field of behavioral economics, anomalies such as self-control problems (present-biased preferences), and susceptibility to how decisions are presented or framed. These anomalies are clearly part of second-best analysis because they imply that consumers are not utility maximizers (and firms may not be profit maximizers) in certain situations. Utility maximization (and profit maximization) in market transactions is a central assumption of the first-best theory. The subfield of behavioral public finance addresses the implications of these behavioral anomalies in the analysis of public policies. A central issue is determining how the government can either undo particular anomalies or exploit them to generate outcomes that promote social welfare.

REFERENCES

Atkinson, A., Stern, N., January 1974. Pigou, taxation, and public goods. Review of Economic Studies 41 (1), 119–128.

Boadway, R., July 1975. Cost–Benefit rules and general equilibrium. Review of Economic Studies 42 (3), 361–374.

Boadway, R., November 1976. Integrating equity and efficiency in applied welfare economics. Quarterly Journal of Economics 90 (4), 541–556.

Boiteux, M., September 1971. On the management of public monopolies subject to budgetary constraints. The Journal of Economic Theory

3. If two or more services commonly use one or more resources, it is not always possible to define average costs for each of them, even in the long run. See Boiteux (1971).

English article 3 (3), 219–240 (translated from French in Econometrica, January 1956).

Diamond, P.A., November 1975. A many person Ramsey tax rule. Journal of Public Economics 4 (4), 335–342.

Diamond, P.A., Mirrlees, J., March, June 1971. Optimal taxation and public production (2 parts; Part I: Production Efficiency, Part II: Tax Rules). American Economic Review 61 (1), 8–27; 61 (3), 261–278.

Hartwick, J., February 1978. Optimal price discrimination. Journal of Public Economics 9 (1), 83–89.

Leuthold, J., February 1976. The optimal congestion charge when equity matters. Economica 43 (169), 77–82.

Pigou, A.C., 1947. A Study in Public Finance, third ed. Macmillan, London.

Wildasin, D., April 1984. On public good provision with distortionary taxation. Economic Inquiry 22 (2), 227–243.

Chapter 19

Transfer Payments and Private Information

Chapter Outline

Our second-best analysis of transfer payments began with the discussion of expenditure incidence in Chapter 17. We noted there that transfers are fully equivalent to negative taxes when the distortions take the form of distortions in prices from their first-best values. The equivalence of taxes and transfers is more general. Because transfers are just negative taxes, taxes and transfers are, in principle, analytically equivalent except for the sign in any policy environment. Nonetheless, transfer payments raise a different set of practical issues from taxes in an environment made second best because of private or asymmetric information.

The practical differences arise because the government's mechanism design problem differs for taxes and transfers under private information. The problem for taxes is to prevent people from evading their tax liabilities. The problem for transfers is to prevent people from accepting transfers that are not meant for them. To this end, the analysis of transfer payments under private information has yielded a number of insights on three issues: whether transfers should be cash or in kind, whether transfers can be decentralized through market-based subsidies rather than being provided directly by government agencies, and the limits of redistribution in the presence of private information. Chapter 19 discusses these three issues, after considering a number of other practical design issues related to cash transfers.

FIRST-BEST INSIGHTS

A quick review of what first-best analysis has to say about transfer payments will be useful as a starting point. The two main decision rules for first-best transfers are the interpersonal equity (IE) conditions of social welfare maximization and pareto-optimal redistributions. The IE conditions are always applicable in the mainstream public sector model since they are among the necessary conditions for a social welfare maximum. The pareto-optimal conditions are applicable if people are altruistic.

IE Conditions

The IE conditions are indifferent between cash or in-kind transfers or taxes. The only requirement is that the transfers (taxes) be lump sum. The operative principle is that if any one good or factor is redistributed to satisfy the IE conditions and the pareto-optimal conditions for a social welfare maximum hold, then the IE conditions are

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00019-4

necessarily satisfied for all the other goods and factors (a cash redistribution can be thought of as a redistribution of income earned by a factor in fixed supply).

The lump-sum redistributions cannot be decentralized; they must be undertaken by the government. Only the pareto-optimal conditions can be decentralized under first-best social welfare maximization, and then only in the absence of externalities, decreasing-cost production, monopoly power, and other such problems that may require more direct government intervention.

Finally, the range of possible redistributions through lump-sum taxes and transfers is as broad as possible, limited only by the boundaries of the first-best utility—possibilities frontier.

Pareto-Optimal Redistributions

Both the form of the transfers and the ability to decentralize depend on the nature of people's altruistic impulses. Think in terms of the rich being altruistic toward the poor.

One possibility is that the utility of the rich depends on some item(s) of consumption by the poor, such as the amount of food they have. In this case the pareto-optimal conditions call for in-kind transfers and decentralization. The poor's purchases of food are directly subsidized and the poor can buy as much food as they want at the subsidized price. The subsidy equals the sum of each rich person's MRS between the poor's consumption of food and his or her consumption of the numeraire good. The only caveat is if the poor's purchases are restricted for some reason, such as to discourage resales of the subsidized item by the poor. The food subsidy is equivalent to a cash transfer if the restriction is binding.

The other possibility is that the utility of the rich depends on the entire utility of the poor, that is, on the amount of resources the poor have. In this case, the transfers should be cash because a cash transfer maximizes the utility of the poor for a given dollar amount transferred. The cash transfers (and taxes on the rich) have to be centralized.

Finally, pareto-optimal redistributions are gain—gain propositions—both the donors and the recipients gain. As such, they restrict the range of the utility—possibilities frontier to the region along which all gain—gain redistributions have been exhausted. The pareto-optimal conditions cannot determine the optimal position on the restricted frontier, however. In the social-welfare-maximizing context of the mainstream model, the IE conditions must be applied as a second layer of redistribution to reach Bator's bliss point on the restricted frontier.

The Samaritan's Dilemma

A dynamic variation of the pareto-optimal redistribution model that we did not discuss in Chapter 10 is the *Samaritan's dilemma*, a phrase coined by James Buchanan. The dilemma can arise whenever the donor and recipient interact for at least two periods. The central feature of the dynamic model is moral hazard on the part of the recipient who, having received a transfer in the first period, can influence the probability of receiving transfers in subsequent periods by relying on the goodwill of the donor. The model calls for an in-kind transfer in the first period to avoid the moral hazard.

Think of parents and their teenaged son. Suppose the parents are purely altruistic toward their son and give him a large amount of cash. The son has an incentive to squander the cash on consumption goods rather than put it to some productive use such as education if he is confident that his parents will never allow him to suffer. The son knows that the parents are "Samaritans" who will provide more cash transfers in the future, even though he has behaved badly in the past.

The parents' dilemma is that they want what is best for their son. On the one hand, they know that the son would prefer a cash transfer that he can spend as he wishes. On the other hand, they can override the moral hazard incentive if they tie the gift first period to a productive endeavor, such as paying the son's tuition for a college education. By tying the aid in this way, the parents prevent the son from using the gift inefficiently and they also reduce or eliminate the need for future gifts.

In fact, a large percentage of *inter vivos* giving from parents to their children is in kind rather than cash. One explanation for this is simple paternalism, that is, parents think they know what is best for their children and they control the resources. Another possibility, however, is simply the desire to avoid the moral hazard associated with a gift of cash. Hence, the dynamic model suggests that in-kind transfers may be efficient even if altruistic donors consider the entire utility of the recipients, whereas the static model would call for cash transfers in that case.[1]

CASH TRANSFERS: BROAD BASED OR TARGETED?

Any society that chooses to redistribute income (cash) to alleviate poverty faces an immediate practical difficulty. It must decide between a broad-based or targeted approach to redistributing, and each has its strengths and weaknesses.

The simplest broad-based approach is to use a so-called *credit income tax* of the form $\text{Tax} = -C + T(Y)$. Everyone receives a credit of $\$C$ and then pays tax on their entire income according to the function $T(Y)$. The credit is refundable, meaning that it is received even if $T(Y) < C$.

1. Buchanan, 1975, pp. 71—85. For a more recent discussion, see Bruce and Waldman, 1991, pp. 1345—1351.

The targeted approach offers subsidies only to people with low incomes and also exempts low levels of income from taxation under the income tax. This is the approach chosen by the United States. It targets subsidies to the poor under various public assistance programs: Temporary Assistance to Needy Families (TANF), Supplemental Security Income, Medicaid, Food Stamps, the Earned Income Tax Credit (EITC), and a number of smaller programs. Then the federal and state income taxes exempt the first dollars of income from taxation through a combination of personal exemptions and standard deductions, such that taxpayers with incomes below the poverty line usually have no tax liability.

A simple example offered by Edgar and Jacquelene Browning illustrates the trade-offs between the broad-based and targeted approaches (Browning and Browning, 1983). Suppose there are five income classes, with equal numbers of people in each income class: $5000, $10,000, $15,000, $20,000, and $25,000. Suppose also that society wants to distribute $3000 to the lowest income group. Two policies that accomplish this are

1. A (linear) credit income tax, $T = -\$4500 + 0.3Y$.
2. A $3000 subsidy to the lowest income class, paid for by a (linear) income tax that exempts the first $10,000 of income and taxes income above $10,000 at a flat rate of 10%.

Table 19.1 provides the subsidies and taxes paid by each income class under these alternatives.

The credit income tax leads to a lot of churning. It collects and distributes $22,500, with a net redistribution of only $4500. It also requires a fairly high marginal tax rate of 30%, which could generate substantial deadweight losses. The example shows that the credit income tax is not a very effective redistributive mechanism. Even a modest amount of redistribution to the poor can require a very large tax-transfer program that may generate a large amount of deadweight loss.

The targeted approach appears to be much more effective on the surface. It collects and transfers only $3000 and the marginal tax rate on the taxpayers is only one-third as high as the credit income tax, leading to much less deadweight loss (recall that deadweight loss increases with the square of the tax rate). The drawback, however, is that the low-income people who are subsidized face an extremely high marginal tax rate. A person at $5000 who works hard and increases her income to $10,000 is only $2000 better off ($8000 versus $10,000). The loss of the $3000 subsidy is in effect a 60% marginal *tax* on the additional $5000 of earnings. A targeted approach thus places a society in the uncomfortable position of expecting the able-bodied poor to work their way out of poverty rather than accept a public subsidy while at the same time subjecting them to very high marginal tax rates. The marginal tax rates can be enormous with multiple targeted public assistance programs such as those in the United States. Some poor families receive aid under five or six different programs. If they work their way out of poverty they not only face a high marginal tax rate because they lose their monthly cash subsidy but also may lose access to Food Stamps, Medicaid, housing assistance, and other subsidies so that their combined marginal tax rate is well in excess of 100%. Indeed, the marginal rates can be so high for some families that they are actually worse off if they earn additional income, in violation of Feldstein's no-reversals principle of vertical equity.

The potential for reversals under targeted assistance is referred to as the *notch problem*. Notch problems always arise at the cutoff points at which the assistance ends. Taxes levied on incomes immediately above the notch are likely to make those taxpayers worse off than people with incomes just before the notch who are still receiving

TABLE 19.1 Subsidies and Taxes Paid by Income Class

	Income Class				
	$5000	**$10,000**	**$15,000**	**$20,000**	**$25,000**
Credit income tax (T = −C + 0.3Y)					
Subsidy ($)	4500	4500	4500	4500	4500
Tax ($)	1500	3000	4500	6000	7500
Net tax (subsidy) ($)	(3000)	(1500)	0	1500	3000
Targeted subsidy ($3000) (T = 0.1(Y−10,000))					
Subsidy ($)	3000	0	0	0	0
Tax($)	0	0	500	1000	1500
Net tax (subsidy) ($)	(3000)	0	500	1000	1500

subsidies. The only way to avoid the notch problem near the cutoff is to decrease the subsidies as incomes approach the notch and exempt a range of incomes above the notch from tax. The example, though discrete, illustrates this principle. The subsidies stop at $5000, whereas the taxes begin at $15,000. If incomes were continuous rather than discrete, the notch problem would arise from $5001 to $7999. People with incomes in that range would be worse off than people with incomes of $5000.

An Acceptable Public Assistance Program?

High marginal tax rates and notch problems are not specific to this example. They are inherent in all targeted transfer programs, which helps to explain why people in the United States and other countries never seem to be satisfied with their public assistance programs. In our view, targeted public assistance programs can never be entirely acceptable. The problem is twofold. On the one hand, we believe that people will be satisfied with public assistance only if it satisfies three goals:

1. It removes virtually everyone from poverty (the poverty gap, the aggregate amount of income by which the poor fall below the poverty line, is approximately $200 billion in the United States (2014)).
2. It is not too costly to the taxpayers.
3. It preserves incentives to work (and maintain families intact).

On the other hand, the only sensible way to design an income subsidy is to define a cutoff level of income, Y_{cutoff}, below which the family (individual) is subsidized, and then set the subsidy equal to some proportion of the difference between the cutoff and actual levels of income.[2] That is,

$$S = x\left(Y_{\text{cutoff}} - Y_{\text{actual}}\right), \quad \text{where } Y_{\text{actual}} \leq Y_{\text{cutoff}} \tag{19.1}$$

with

$$Y_{\text{total}} = Y_{\text{actual}} + S \tag{19.2}$$

The proportion x has to be fairly large, between 0.5 and 1, to generate a large enough subsidy for people with very low incomes. Subsidies of this form have a guaranteed minimum level of income equal to xY_{cutoff}, the subsidy for a family with no income. The subsidy decreases by xY_{actual} as the income increases up to Y_{cutoff}, when it becomes zero. All the cash public assistance programs except the EITC (see below) take this form.

The problem is that a subsidy of this form cannot satisfy all three goals. Suppose Y_{cutoff} is set equal to the poverty line, approximately $24,000 for a family of four in the United States (2014). The program will not be "too costly," in line with the second goal, but it miserably fails the first goal—no one escapes poverty. Everyone below the poverty line remains below the poverty line after receiving the subsidy. To satisfy the first goal, Y_{cutoff} has to be set at $Y_{\text{poverty line}}/x$, so that the guaranteed minimum income equals the poverty line and everyone with $Y_{\text{actual}} > 0$ ends up with Y_{total} above the poverty line. This could be very costly in terms of lost tax revenues, however, if x is much less than 1. Many families with incomes above the poverty line would be subsidized rather than taxed, and the first dollars of taxable incomes above Y_{cutoff} must be exempt from tax to avoid the notch problem.

Regardless of the trade-offs between the first two goals, the third goal is impossible to achieve with a subsidy of this form. It has the strongest possible disincentives for work (and maintaining families intact, for that matter).[3] A person who earns extra income in the subsidized range below the cutoff still receives a subsidy: $Y_{\text{total}} > Y_{\text{actual}}$. Hence, there is an income effect from the total subsidy that favors leisure over work. At the same time, additional earned income is subject to a marginal tax rate of $x\%$ because of the loss of subsidy at the rate of x per dollar. Hence, there is a substitution effect that again favors leisure over work. The combination of a subsidy on average and tax on the margin is doubly destructive for work incentives.

The federal government tried to avoid the work incentive problem when it established three public assistance programs as part of the Social Security Act of 1935. It chose a categorical approach, giving aid only to those who were deemed to have well-below-average prospects for work and were therefore not expected to lift themselves out of poverty. Thus, the first three programs targeted aid only to the elderly, the blind, and single-parent families (primarily widows, who usually had little or no insurance and thus were in desperate straits if their husbands died). Aid to the disabled was added in 1951. An obvious drawback to categorical targeting is that poor families who do not fall into one of the four categories receive no aid at all, such as the majority of poor two-parent families. The United States apparently felt in 1935, and long afterward, that keeping large numbers of families out of the public assistance safety net was an acceptable price to pay to avoid incentive problems.

Incentive problems appeared anyway in the 1960s when the single-parent program, Aid to Families with Dependent Children, exploded. The nuclear family was weakening, and single women with children began seeking public assistance because their husbands had deserted them, not because they had died. Taxpayers felt cheated for the first time and searched for ways to improve incentives to work and to keep low-income families together.

2. Y_{cutoff} would also vary with family size.

3. The family has a strong financial incentive for one spouse to leave home and send back income in a manner that cannot be detected, thereby avoiding the high marginal tax rates built into the public assistance formulas.

The EITC

The best one can do in terms of work incentives is a wage subsidy of the form:

$$S = xY_{\text{actual}} \tag{19.3}$$

Unfortunately, the income effect remains in effect and favors leisure over work. But at least the substitution effect favors work, since additional earnings are subsidized at rate x rather than being taxed at rate x. The basic public assistance programs cannot take this form, however, because then people with no income starve to death. But a supplemental antipoverty program could take this form to alleviate the high marginal tax rates of the basic program(s).

This is exactly what the EITC was designed to do. As noted above, the combined marginal tax rates under the other public assistance programs were enormous for some of the poor. The EITC offset this somewhat by offering wage subsidies to the poor that varied by family size. In 2014, for example, a single parent with two children received a 40% subsidy on wage and salary income to an income of $13,400, a maximum subsidy of $5372 at the cutoff. Notice, however, that the subsidy cannot simply end without generating an enormous notch problem, an inherent drawback with wage subsidies. Thus, the subsidy remained fixed at $5372 for incomes between $13,401 and $17,750. Beyond, $17,750, the subsidy is phased out at the marginal rate of 21 cents on the dollar until it reaches zero, at $43,000. The income ranges are increased each year by the rate of inflation.

The EITC became a very large program in the 1990s, reaching $26 billion by financial year 2000 (by comparison, TANF was $30 billion that year). It helped considerably to reduce the marginal tax rates on poor families and thereby encourage their work effort. The EITC is not without its drawbacks, however. The program discourages work effort in the flat subsidy range, given the income effect, and doubly discourages work effort in the phase-out range. The marginal tax rate increases by 21 percentage points in the phase-out range. The combination of the 15% and 25% income tax rates in the phase-out range (the rate jumps from 15% to 25% at $36,300), the 21% phase-out rate, the 15+% payroll tax rate (assuming labor bears the entire burden), and state income tax rates (in 44 states) saddle many low-income families with marginal tax rates well in excess of 50%. Also, the vast majority of the subsidies under the EITC go to the nonpoor. In summary, the need to avoid the notch problem under a wage subsidy sharply reduces its attractiveness as a means of fighting poverty.[4]

The United States decided during the Reagan administration to give up on trying to preserve work incentives under the basic public assistance program benefit formulas. The subsidy proportion x was set equal to 1 on all income after the first four months, thereby completely destroying any incentive to work below Y_{cutoff}. Families receive Y_{cutoff} no matter what their actual incomes are. Setting $x = 1$ does well by the first two goals, so long as Y_{cutoff} is the poverty line.[5] But it requires a stick approach to maintain the work incentive called *workfare*, in which the single parent is forced to work (or receive education or job training) in order to receive benefits. Workfare was applied hesitantly by some of the states until 1996, when it became the cornerstone of the TANF program that replaced Aid to Families with Dependent Children. TANF removed public assistance as an entitlement. States were allowed to remove families from the welfare rolls after 2 years of receiving benefits. They were also encouraged to force able-bodied welfare parents to undertake job training or work in order to maintain their benefits during the first 2 years. The federal government also gave financial support to the states to provide child support for those who entered the workforce or training programs, and expanded Medicaid so that working mothers with children would not lose medical care as their incomes increased.

The combination of the work requirements under TANF, the EITC, and the expansion of Medicaid proved very successful in encouraging single parents to leave public assistance for work. The number of families receiving TANF assistance fell by 66% from 1994 to 2007, when the US economy remained close to full employment. But then the number of families receiving TANF increased by 8% from 2007 to 2010 when the Great Recession hit (Falk, 2013). These numbers illustrate one problem with the stick approach, that the ability of people on public assistance to work depends very much on the state of the economy, the extent to which jobs are available to them. Also, workfare under TANF is not entirely comforting since the stick is being applied primarily to women with very low incomes who are trying to raise their children on their own. The work incentive problem associated with targeted income subsidies to the poor has no obvious solution.

Special Needs, In-Kind Transfers, and Universality

Suppose the government decides to offer in-kind transfers to pay for special needs, such as insulin shots for diabetics, eyeglasses, or hearing aids. Nicholas Rowe and Frances Wooley have argued that these transfers should probably be universal—offered to everyone with these special

4. For further discussion and analysis of the EITC, see Browning, 1995; Schloz, 1994.

5. In fact, Y_{cutoff} is set below the poverty line under TANF and Supplemental Security Income in all states, and often well below the poverty line.

needs—rather than targeted to the poor (Rowe and Wooley, 1999). The case for universal transfers follows from thinking of them in the context of the Mirrlees optimal income tax (transfer) problem with utilitarian social welfare that we discussed in Chapter 15.

Suppose the needs are observable, so that the only information problem in the optimal income tax framework remains the inability to know people's skill level. Suppose further that a special need reduces effective consumption dollar-for-dollar by the expenditures required to offset the need and does so equally for everyone. Also, the special needs are independent of the skill level, as is likely for the examples given above and many other special needs. Under these assumptions, the common utility function in the optimal income tax problem is U^{ij} $(C_{ij}-N_i, 1-L_{ij})$,where:

C_{ij} = consumption of person j with special need i,
N_i = expenditures required to address special need i, equal for everyone with the special need, and
$1-L_{ij}$ = labor supplied by person j with special need i.

Define the "poor" as everyone below a given skill level and the "rich" as everyone at or above that skill level. Suppose in-kind transfers equal to N_i are targeted to the needy poor, with a phase-out extending into the nonpoor range. Targeting in this way is suboptimal in two respects in the context of optimal income taxation. First, the unneedy rich have higher effective consumption than the needy rich at each skill level and, therefore, a lower marginal utility of consumption. Social welfare would be increased if their marginal utilities of consumption were equalized under utilitarian social welfare. This is the inequity of targeting. Second, the loss of benefits in the phase-out range causes the needy and unneedy in that skill range to face different effective marginal tax rates. Assuming they have equal compensated labor supply elasticities, they should face the same marginal tax rates. That some people of equal skills face unequal marginal tax rates is the inefficiency of targeting.

Both the inequity and inefficiency of targeting are avoided if the tax-transfer function has the universal form $T_i = f\ (Y_{ij}) - N_i$. That is, the in-kind transfer is given to everyone who is needy. Rowe and Wooley define universality in the presence of special needs as a tax-transfer function that is additively separable in income and needs.

One obvious condition under which universality would be suboptimal in the context of the optimal income tax problem is if the compensated labor supply elasticities were correlated with the special need. This may apply for certain kinds of disabilities, but probably not for many other kinds of special needs such as those given above. In any event, Rowe and Wooley argue that universality should be the benchmark for special needs transfers against which exceptions have to be justified.

Many people would argue that targeting in-kind transfers to the poor is equitable in the sense that it transfers purchasing power from the rich to the poor, but this argument does not apply to special needs. Here the potential for inequity is between the unneedy and the needy rich, which targeting gives rise to and universality avoids. The optimal income tax does transfer purchasing power from the rich to the poor, but Rowe and Wooley's point is that in-kind transfers to offset special needs should probably not be part of the rich-to-poor transfer of purchasing power.

PRIVATE INFORMATION AND IN-KIND TRANSFERS

The federal government has always enhanced its cash public assistance transfers with in-kind subsidies to the poor, primarily for medical care, food, and housing assistance. The original three cash assistance programs, and later Aid to the Disabled, included medical vendor payments—that is, payments to physicians and hospitals that provided medical care to the recipients. The medical assistance was consolidated into a single program, Medicaid, in 1965. Medicaid grew rapidly in the 1970s and the 1980s, in part because medical costs experienced high inflation and in part because Medicaid was expanded to cover the so-called medically needy families. These are families with low incomes who are not on public assistance but who have large medical expenses. Medicaid was considerably expanded in the 1990s, primarily to cover children and pregnant women in families with incomes as much as twice the poverty line. Housing assistance is another in-kind program that has long subsidized poor families. Food Stamps were added in the early 1970s. The in-kind subsidies are now much larger than the cash subsidies. Spending under Medicaid alone is approximately $450 billion per year (2014), more than all the other public assistance programs combined, cash and in kind.

Recall that the first-best pareto-optimal redistribution model provides a justification for in-kind aid and suggests that it should be decentralized. The government should ideally subsidize the poor's purchases of food, housing, or medical care and let them buy as much as they wish in the private market at the subsidized rate. This is presumably what the citizens in a capitalist country would want the government to do. They would not want government agencies providing these goods using some kind of nonmarket rationing device.

The Blackorby–Donaldson Model of In-Kind Transfers

Unfortunately, the decentralized subsidy approach becomes vulnerable in the presence of private information. Charles Blackorby and David Donaldson developed a very simple

model of the provision of medical care that shows that government provision (rationing) is quite likely to be preferred to decentralized subsidies if the government cannot be certain who really needs medical care.[6] We saw in Chapter 15 how private information limits the government's tax options. The limitations on transfers imposed by private information can be even more severe. The Blackorby–Donaldson model clearly demonstrates why private information limits the government's ability to target transfers to the needy no matter what form they take, cash or in kind, and if in-kind whether by rationing or by subsidy.

The model consists of two classes of individuals. One class, H, is healthy, so that the utility of class H individuals is a function only of their income (consumption of private goods—in the model, a numeraire composite commodity):

$$U_H = Y_H \tag{19.4}$$

The second class, I, has an illness, the disutility of which can be reduced by medical services Z; therefore, the utility of class I individuals depends on both their income (consumption of private goods) Y_I and Z according to the utility function:

$$U_I = Y_I - e^{(1-Z)} \tag{19.5}$$

Individuals within each class are identical.

The individuals possess private information about the state of their health. The government knows that there are two classes of consumers and knows the utility functions for each class, but it cannot know whether any one individual is healthy or ill. Since the individuals are identical within each class, we will consider the simplest case of a two-person economy with one H person and one I person.

The economy is endowed with 6 units of a resource, K, and a unit of K can be transferred into either a unit of private goods or a unit of Z. Therefore, the production–possibilities frontier for the economy is

$$Y_H + Y_I + Z = 6$$

The First-Best Frontier

The first task is to establish the first-best utility–possibilities frontier, which the government could achieve if it had perfect information about individuals' health. The first-best frontier serves as a baseline for comparing the rationed and subsidy approaches to the provision of Z under private information.

The necessary condition for the economy to be on its first-best frontier is that $\text{MRS}_{Y,Z} = \text{MRT}_{Y,Z}$. The $\text{MRT}_{Y,Z} = 1$, and the $\text{MRS}_{Y,Z}$ applies only to those who are ill. Thus, first-best pareto optimality requires that:

$$\text{MRT}_{Y,Z} = \partial U_I / \partial Z \big/ \partial U_I / \partial Y_I = e^{(1-Z)} = 1 \tag{19.6}$$

or that $Z = 1$.

The first-best utility–possibilities frontier is pictured in Fig. 19.1. At one extreme I gets all the resources, purchases 1 unit of Z and 5 units of Y_I, and has $U_I = 4$. Given that $Z = 1$, the other extreme consists of H receiving 5 units of Y_H, in which case $U_H = 5$, and $U_I = -1$. The region from $U_H = 5$ to $U_H = 6$ implies that Z is less than 1. I prefers Z to Y in that region because Z has the higher marginal utility with $Z < 1$. Therefore, U_I falls along the curved line to the limit of $-e$, when I has no resources.

Now introduce the private information and consider two government policies: (1) government provision—rationing—of the medical care and (2) subsidizing the purchase of medical care by I, the decentralized solution.

Government Provision of Medical Care

The most straightforward way to consider government provision in this simple model is to assume that the government allocates Y as well as I. Therefore, it has complete control to place the economy anywhere on the second-best utility–possibilities frontier under private information.

Private information gives rise to the mechanism design problem of ensuring that the medical care is given only to those who are ill. The self-selection constraints are

$$Y_H \geq Y_I \text{ for person } H \tag{19.7}$$

$$U_I(Y_I, Z) \geq U_I(Y_H, 0) \text{ for person } I \tag{19.8}$$

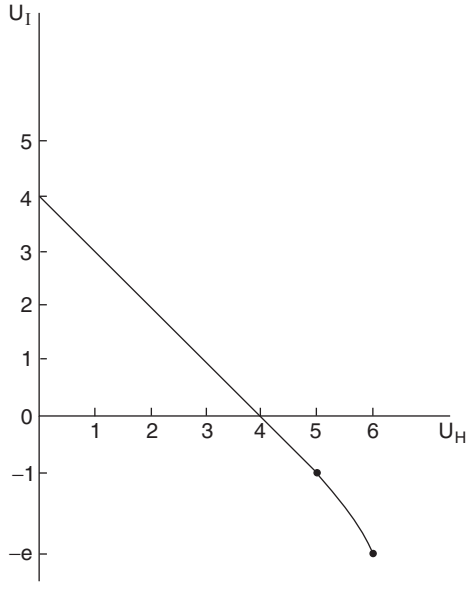

FIGURE 19.1

6. Blackorby and Donaldson, 1988. Note the difference relative to Rowe and Wooley, who assume that the need for medical care is known to the government.

H has no use for Z; therefore, she will identify herself correctly as long as the government gives more Y to the healthy. I will identify himself correctly unless H receives so much more Y that it pays for I to forego Z and declare himself as healthy. Equation (19.8) gives the combinations of Y_I and Z for which he will correctly identify himself as ill.

The first point to note is that some of the first-best frontier is preserved under government provision. Z must equal 1 to be on the first-best frontier. This leaves 5 units of Y to be distributed between H and I. Equation (19.7) sets a lower bound for Y_H equal to 2.5, since H must always have at least half of the total Y. Equation (19.8) sets the upper bound of Y_H as follows:

$$Y_I - 1 \geq Y_H - e \tag{19.9}$$

$$(5 - Y_H) - 1 \geq Y_H - e \tag{19.10}$$

$$Y_H \leq (2 + e/2) \tag{19.11}$$

$$Y_H \leq 3.4(\text{approximately}) \tag{19.12}$$

Refer to Fig. 19.2. Within the region $2.5 \leq Y_H \leq 3.4$, the second-best utility–possibilities frontier with private information coincides with the first-best frontier. Outside that region, the frontier with private information must be below the first-best frontier. If $Y_H < 2.5$, then Y_I must also be <2.5 to satisfy the self-selection constraint Eqn (19.7). But this implies $Z > 1$, in violation of the first-best pareto-optimal condition, Eqn (19.6). At the upper boundary of $Y_A = 3.4$, I has 2.6 units to divide between Y_I and Z. The division $Y_I = 1.6$, $Z = 1$, yields the same utility (approximately) for I that he would have by declaring himself healthy and taking 3.4 units of Y with no Z:

$$U_I = 1.6 - 1 = .6 = 3.4 - e \tag{19.13}$$

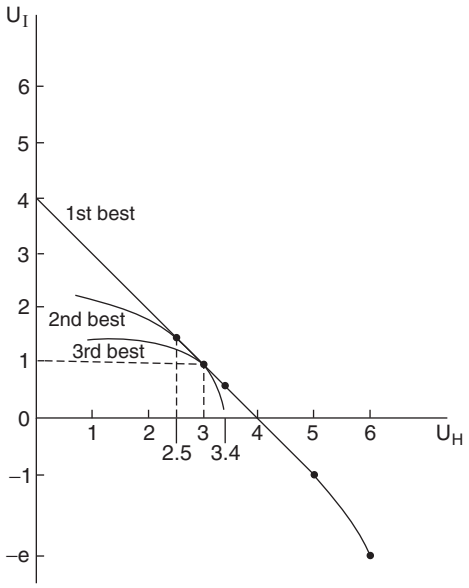

FIGURE 19.2

If the government tried to set $Y_H > 3.4$, then the ill would declare themselves healthy. There is no combination of Z and Y_I with total resources less than 2.6 that yields them as much utility as taking Y_H, since Z and Y are both normal goods. But society does not have enough resources to give both people 3.4 units of Y, so the second-best frontier stops at $Y_H = 3.4$.

Subsidizing Medical Care

Private information severely constrains the decentralized subsidy approach. With the government unable to determine who is healthy and who is ill, the individuals will always choose the option with the highest purchasing power and use it as they wish. Therefore, the government has no choice but to equalize purchasing power under the subsidy plan. The single self-selection constraint is

$$Y_H = Y_I + qZ \tag{19.14}$$

where q is the subsidized price of Z.

Refer again to Fig. 19.2. The only attainable point on the first-best frontier is equal resources with $q = 1$. Setting the subsidy to zero ($q = 1$) is necessary to satisfy the first-best pareto-optimal condition, Eqn (19.6); equalizing the resources is necessary to satisfy the self-selection constraint, Eqn (19.14). Thus, the only feasible first-best allocation is

$$\{Y_H = 3; Y_I = 2, Z = 1; U_A = 3; U_I = 1\}.$$

Subsidizing medical care ($q < 1$) favors I, and taxing medical care ($q > 1$) favors the H with Eqn (19.14) holding, but in either case society is below the first-best frontier as shown in the figure. Furthermore, the frontier with subsidy must be below the second-best frontier with rationing in the subsidy region. It is clearly below the rationing frontier above $Y_H = 2.5$, since the rationing frontier is the first-best frontier in that region. It is also below the second-best rationing frontier when $Y_H < 2.5$ since it adds the binding equal-purchasing-power constraint that is absent under rationing.

In conclusion, the Blackorby–Donaldson model demonstrates that private information generates a preference for rationing (government provision) over decentralized subsidies when society's charitable impulse is to give people in-kind aid. The intuition is that rationing prevents people from claiming a subsidy to treat some illness that they do not have. Rationing may be the only way of verifying the illness. This result overturns the conclusion of the first-best model of pareto-optimal redistribution, which calls for decentralized subsidies when the charitable impulse is for in-kind aid. Even so, rationing preserves some of the first-best allocations.

The model also points to a more general problem that private information poses for the mainstream public sector

model in its attempt to achieve end-results equity. Suppose the government tries to satisfy the first-best IE conditions with lump-sum cash subsidies to the poor. If the government cannot know who is poor and who is nonpoor, then everyone will claim to be poor in order to receive a subsidy. The only way to prevent people from hiding their true identities is to impose equal incomes for everyone. No other income redistribution is feasible.

We saw in Chapter 4 that the first-best model does imply complete equality under the three assumptions of equal social welfare weights (at equal utilities), identical tastes, and diminishing marginal utility of income. The private information result is stronger, however, since equal income is the only feasible outcome regardless of the assumptions one chooses to make about the social welfare function, tastes, and marginal utility. People's willingness to exploit their private information can clearly be devastating to the mainstream ideal of the government sector acting as an agent for the citizens in the pursuit of the public interest in efficiency and equity. Remember that if the first-best IE conditions cannot be satisfied, then in general the first-best pareto-optimal conditions are not optimal either. Also, second-best interventions are generally all-pervasive rather than limited in scope, not at all in the spirit of the government-as-agent ideal.[7]

The Besley–Coate Model of Workfare

As discussed above, workfare was adopted as a centerpiece of TANF to force welfare parents to prepare for and accept jobs in an attempt to overcome the strong work disincentives of the benefit formula. Workfare responds to a principle of long standing in the United States, that the able-bodied should work rather than simply accept a handout from the government. It also tries to help welfare parents become self sufficient by improving their prospects in the labor market.

Timothy Besley and Stephen Coate have shown that workfare can be useful even if the work enforced by the government is entirely unproductive to society or the individual. In particular, unproductive workfare can serve as a signaling device that allows the government to target cash subsidies to the poor in a world of private information in which it would otherwise be difficult to distinguish the poor from the nonpoor. Besley and Coate were responding to models such as that of Blackordy and Donaldson's, which concludes that targeting cash subsidies to the poor is impossible if the government cannot distinguish among individuals. Its only option is to equalize incomes. In the Besley–Coate model, workfare acts as a self-selection mechanism that prevents the nonpoor from accepting the public assistance subsidies.[8]

The Besley–Coate model has the following elements:

Individuals

There are two classes of individuals, those with high ability (H) and those with low ability (L). The high-ability individuals receive a wage of W_H, and the low-ability individuals a wage of W_L. All individuals have the same additive separable utility functions:

$$U_i = Y_i - h(l_i), \quad \text{for } i = H, L \tag{19.15}$$

where Y_i is the numeraire composite commodity (income) and l_i is the labor. The proportions of low- and high-ability individuals are γ and $(1-\gamma)$. Labor markets are competitive, so that $W_H = h'(l_H)$ and $W_L = h'(l_L)$. Letting $\hat{l}_i$ represent the equilibrium labor supplies,

$$Y_L = W_L\hat{l}_L = Y_L(0, W_L) \tag{19.16}$$

and

$$Y_H = W_H\hat{l}_H = Y_H(0, W_H) \tag{19.17}$$

The zeros in Y( ) indicate the absence of government transfers.

The Government

The government has a Rawlsian social welfare function. It wants to ensure that everyone has at least a minimum acceptable level of income Z, with $Z > W_L\hat{l}_L$ (and $Z < W_H\hat{l}_H$). It considers two public assistance plans:

1. A straight welfare plan with lump-sum subsidies b_L and b_H targeted to the low- and high-ability individuals, respectively.
2. Workfare, which includes a forced work requirement of C units of labor in order to receive the lump-sum subsidies b_L or b_H.

7. P. Bearse et al. offer another possible explanation for the well-documented preference for in-kind aid in less developed countries: the difficulties these governments have in raising tax revenues. They present a simple model in which people have preferences over a composite commodity and education. The government raises taxes to provide either a universal transfer or universal public education, free to all. Families can opt out of public education in favor of private education. People vote in a direct democracy for the level of the tax rate and also for the shares of tax revenues devoted to the transfer and to public education. In their model, as taxes become more difficult to collect the proportion of revenues devoted to public education increases. The main reason is that with lower tax collections the quality of public education suffers and more of the higher income people opt out for private education. The opting out of public education by the rich increases the return of tax dollars spent on education relative to the universal transfer from the point of view of the poor. This leads the majority of voters, who are poor, to prefer an increase in the share of tax revenues devoted to public education (see Bearse et al., 2000).

8. Besley and Coate, 1992b. They extended and generalized the analysis in Besley and Coate, 1995.

The enforced work is entirely unproductive. Its only purpose is to serve as a self-selection device. Also, workfare has no effect on the total amount of labor supplied by either type of person because utility is additively separable. C substitutes one-for-one for market work, so that incomes under workfare are

$$Y_L = W_L(\hat{l}_L - C) = Y_L(C, W_L) \tag{19.18}$$

and

$$Y_H = W_H(\hat{l}_L - C) = Y_H(C, W_H) \tag{19.19}$$

Finally, straight welfare and workfare are costless to administer.

The government's goal is to minimize the cost of public assistance subject to satisfying the Rawlsian minimum income constraint:

$$\begin{aligned} &\min\ \gamma b_L + (1-\gamma) b_H \\ &\text{s.t. } Y_i + b_i \geq Z, \quad \text{for } i = L, H \end{aligned}$$

An additional requirement is that participation in straight welfare or workfare has to be voluntary. This requirement is most easily represented in terms of the indirect utility function, with the lump-sum subsidies, workfare, and the market wages as parameters:

$$V_i(b_i, C, W_i) \geq V_i(0, 0, W_i), \quad \text{for } i = L, H \tag{19.20}$$

First-Best Optimum

The cost-minimizing option with perfect information would clearly be a straight welfare program that only subsidizes the low-ability people such that they achieve Z:

$$b_L = Z - W_L\hat{l}_L; b_H = 0; C = 0; \text{with a cost of } \gamma b_L.$$

Workfare cannot be first best because C is unproductive.

Private Information

Besley and Coate consider two degrees of private information:

1. Earnings are unobservable. This is most likely to apply to a less developed country.
2. Earnings are observable, but effort is not, that is, the government knows Y_i and W_i, but not l_i. The inability to observe effort allows a high-ability individual to claim to have low ability by working just hard enough to earn Y_L with a wage of W_H.

Either type of private information introduces two self-selection (incentive compatibility) constraints:

$$V(b_H, C, W_H) \geq V(b_L, C, W_H) \tag{19.21}$$

$$V(b_L, C, W_L) \geq V(b_H, C, W_L) \tag{19.22}$$

Equations (19.21) and (19.22) require that class H and L individuals prefer their own public assistance options to those of the other class. Equation (19.21) is the operative constraint, because society wants to prevent high-ability individuals from accepting public assistance.[9]

Unobservable Earnings

We will demonstrate how workfare can promote the government's goals in the first case since it is the easier one.

Straight Welfare

With earnings unobservable, the government cannot target cash subsidies to those with low ability. If it tried to set $b_L > b_H$ (with b_H likely equal to zero), the high-ability individuals would claim to have low ability and take b_L. Therefore, the only straight welfare policy is equal subsidies to both, in an amount sufficient to bring those with low ability to Z:

$$b_L = b_H = Z - W_L\hat{l}_L$$

Workfare

Low- and high-ability individuals react differently to workfare:

1. *Low-ability individuals*: The low-ability individuals will accept workfare as long as $C < \hat{l}_L$. Their total labor supply is unchanged, so the disutility of working, $h(\hat{l}_L)$, remains the same. Also, their income is increased to Z, which is better than they can do on their own. The required subsidy is

$$b_L = Z - Y_L(C, W_L) = Z - W_L(\hat{l}_L - C) \tag{19.23}$$

Notice that b_L is larger than under straight welfare because the individuals' earnings are reduced by the unproductive workfare requirement.

2. *High-ability individuals:* High-ability individuals may claim to have low ability in order to receive the subsidy. Whether they do or not depends on the size of the subsidy relative to the cost of accepting the workfare requirement.

If they tell the truth and turn down the subsidy, their utility is $U_H = W_H\hat{l}_H - h(\hat{l}_L)$. The government's mechanism design problem is to find a level of workfare, C^*, such

9. The information set is somewhat unrealistic since the government knows individuals' common utility functions, even though it cannot fully distinguish between people with high and low ability.

that the high-ability individuals have an incentive to tell the truth. Their utility under workfare is $U_H = W_H(\hat{l}_H - C) - h(\hat{l}_H) + b_L = W_H(\hat{l}_H - C) - h(\hat{l}_H) + Z - W_L(\hat{l}_L - C)$. Therefore, the level of workfare that just makes the high-ability individuals indifferent to accepting the subsidy is the solution to:

$$W_H\hat{l}_H - h(\hat{l}_H) = W_H(\hat{l}_H - C^*) - h(\hat{l}_H) + Z - W_L(\hat{l}_L - C^*) \tag{19.24}$$

or

$$W_H C^* = Z - W_L(\hat{l}L - C^*) \tag{19.25}$$

The left-hand side of Eqn (19.24) is the cost of claiming to be low ability, the sacrificed income to workfare, and the right-hand side is the benefit b_L under workfare. With $C > C^*$, the workfare subsidy, Eqn (19.23), is effectively targeted only to the low-ability individuals.

Notice that, from Eqn (19.25), $Z - W_L \cdot \hat{l}_L = (W_H - W_L)C^*$ is the subsidy b_L under straight welfare, a subsidy that would have to be given to everyone. Therefore, workfare with a requirement of C^* is the least-cost public assistance strategy if:

$$(W_H - W_L)C^* > \gamma W_H C^* \tag{19.26}$$

or

$$(l - \gamma)W_H > W_L \tag{19.27}$$

Equation (19.27) indicates that workfare is the preferred strategy if either (1) γ is low, so that there are not so many low-ability individuals to target, or (2) $W_H >> W_L$. The lower (relatively) W_L is, the more costly it is to subsidize everyone. Also, the additional subsidy required because of workfare, $W_L\, C^*$, is that much smaller.

The more complex case of observable earnings but unobservable effort, which is more likely for developed countries, yields two additional results of interest. The first is that workfare is less likely to be cost minimizing relative to the earnings-unobservable case. The reason is that the high-ability individuals have to sacrifice much more to receive the subsidy—they have to earn $W_L \cdot \hat{l}L$. The second is that workfare is better if earnings are low because W_L is low rather than because $h(l)$, the disutility of work, is high. A high $h(l)$ would apply to people who are unemployable or who have a number of children at home to care for. For these people it is better to target aid in some other way, such as government provision of medical care as in the Blackorby–Donaldson model or targeting on the basis of observable characteristics (interested readers should consult the Besley and Coate article).

As noted earlier in the chapter, targeting on observable characteristics was the strategy followed by the United States in the Social Security Act of 1935. It targeted cash subsidies only to those poor who were also likely to have poor employment prospects, such as the elderly and widows. The federal government was well aware of the perils of private information when it entered the public assistance fray. Indeed, public assistance was originally allocated entirely to local governments because towns were generally small and local officials were likely to know the poor. Hence, they would know who was truly deserving of aid and who was shirking. They would not face the problems of private information that beset the states and the federal government when trying to help the poor.

The Besley–Coate model is more sanguine than the Blackorby–Donaldson model about the government's ability to use targeted cash transfers in a world of private information. Yet, it has not entirely rescued the mainstream view of the government from the difficulties of private information. That the government has to resort to enforced workfare of the poor to prevent the nonpoor from taking public assistance is unsettling to the notion of the government acting as an agent for the citizens in pursuit of end-results equity. One wonders if the poor are indifferent between market work and workfare as in the Besley–Coate model, especially if they view workfare as unproductive. If they are not indifferent, then the potential least-cost property of workfare is less compelling. In any event, workfare should be an effective deterrent to the nonpoor falsely claiming public assistance benefits, which is the main point of the Besley–Coate model.

Welfare Stigma

Robert Moffitt's 1983 empirical study of public assistance in the United States found strong evidence that welfare recipients suffer from a stigma that reduces their utility from public assistance (Moffitt, 1983). After 9 years, Besley and Coate developed a simple theoretical framework for analyzing welfare stigma (Besley and Coate, 1992a). An important feature of their model is that stigma is endogenous, determined by such things as the level of public assistance payments and the percentage of "undeserving" poor who accept welfare.

Besley and Coate speculated that the stigma of being on welfare may arise in one of two ways:

1. As a form of statistical discrimination, in which the knowledge that some able-bodied poor choose to accept public assistance rather than work leads to a perception among the nonpoor that all welfare recipients are lazy. All the poor who accept public assistance, even those who are deserving, feel stigmatized by this perception.
2. From taxpayer resentment, in which some of the nonpoor simply resent having to pay taxes to support the poor. That some taxpayers feel this way is enough to generate a sense of stigma among the poor.

Besley and Coate's model of welfare stigma can accommodate both types of stigma, and they have different implications. We will present the analysis of statistical discrimination here.

Elements of the Model

The population consists of N individuals, n of whom are poor, and (N − n) nonpoor. The poor are of two kinds: the deserving ("needy") poor, who cannot work and have been targeted to receive public assistance, and the undeserving poor, who are able to work. The proportions of deserving and undeserving poor are γ and $(1-\gamma)$, respectively. The nonpoor have altruistic feelings toward the deserving poor, which that leads them to support a public assistance program, but the intensity of their altruism varies. Their (common) utility function is

$$U = U(C) - \mu\gamma nP(C_n) \tag{19.28}$$

where:

C = the consumption of each nonpoor;
μ = the altruism parameter, which is distributed among the nonpoor according to the continuous distribution function $G(\mu)$;
γn = the number of deserving poor;
C_n = the consumption of a deserving poor person; and
$P(\ )$ = an index of hardship suffered by a deserving poor person, with $P' < 0$.

A welfare payment b is given to all the poor who do not work. The nonpoor can tell who among the poor are working, but they cannot tell whether a welfare recipient is deserving or undeserving. This is the information problem in the model.

Let M = the number of poor who accept public assistance. Then the per-person tax payment by the nonpoor, T, is

$$T = \frac{Mb}{(N-n)} \tag{19.29}$$

The undeserving poor who choose to work earn a wage w and receive utility of

$$U = V(w) - \theta \tag{19.30}$$

where θ indicates the disutility of working, which varies among the poor. It is distributed according to the uniform distribution function from 0 to 1.

Any poor person who accepts public assistance receives utility of

$$U = V(b) - s \tag{19.31}$$

where s is a measure of the stigma they suffer as recipients of public assistance. The deserving poor have to accept this option because they are unable to work and earn w.

The border of indifference for an undeserving poor person to accept public assistance, given s, is the disutility of work $\hat{\theta}$ that solves:

$$V(w) - \hat{\theta} = V(b) - s \tag{19.32}$$

The undeserving poor with $\theta > \hat{\theta}$ choose public assistance. Therefore, the total number of the poor on welfare is:

$$M = n\left[\gamma + (1-\gamma)\left(1 - \hat{\theta}\right)\right] \tag{19.33}$$

The utility of the nonpoor with the public assistance program is

$$U = U\left(C - \frac{Mb}{(N-n)}\right) - \mu\gamma nP(C_n) \tag{19.34}$$

Statistical Discrimination

The statistical discrimination motive for stigmatizing the poor depends on the distribution of the disutility from work among the poor. The average disutility to work among the poor is

$$\overline{\theta} = \int_0^1 \theta d\theta \tag{19.35}$$

which is also assumed to be the average disutility to work among the nonpoor. In other words, $\overline{\theta}$ is the accepted social norm relating to the distaste for work. The average disutility of work among the undeserving poor who choose public assistance is

$$\overline{\theta}_u = \int_{\theta}^1 \theta d\theta / \left(1 - \hat{\theta}\right) \tag{19.36}$$

Therefore, the average disutility of work among all the poor on public assistance is

$$\overline{\theta}_w = \pi\overline{\theta} + (1-\pi)\overline{\theta}_u \tag{19.37}$$

where

$$\pi = \frac{\gamma n}{M} \tag{19.38}$$

which is the proportion of the poor on public assistance who are deserving. Notice that $\pi < 1 \Rightarrow \overline{\theta}_w > \overline{\theta}$.

The stigma based on statistical discrimination arises from the difference between the average disutility of work among the poor, $\overline{\theta}_w$, and the accepted social norm $\overline{\theta}$. Let $g(\overline{\theta}_w - \overline{\theta})$ be the function that generates stigma, such that $g' > 0$ and $g(0) = 0$. $\overline{\theta}_w$ is a function of b and s through the work/accept public assistance relationship, Eqn (19.32).

Therefore,

$$s^* = g\left(\hat{\theta}_w(b, s^*) - \hat{\theta}\right) \tag{19.39}$$

solves for the equilibrium level of stigma given the exogenous variables b and w, the public assistance payment, and the wage. s^* closes the model by determining the number of poor who accept welfare, M, and thus the tax payments of the nonpoor.

An immediate problem is that s^* may not be unique. To see why not, differentiate Eqn (19.39) with respect to s and note that $g'\frac{\partial\overline{\theta}_w}{\partial s}$ must be less than 1 to ensure that s^* is unique. But

$$\frac{\partial\overline{\theta}_w}{\partial s} = (1-\pi)\frac{\partial\overline{\theta}_u}{\partial s} + \left(\overline{\theta} - \overline{\theta}_u\right)\frac{\partial\pi}{\partial s} \tag{19.40}$$

$\overline{\theta} < \overline{\theta}_u$ and $\frac{\partial\pi}{\partial s} > 0$ (the proportion of deserving poor on public assistance increases with stigma as more of the undeserving poor choose to work). Therefore, the second term is negative. The first term is positive because as some undeserving poor choose to work the average disutility from work of the undeserving poor who remain on public assistance rises. Thus, $\frac{\partial\overline{\theta}_w}{\partial s}$ could be greater than 1. Besley and Coate assume that $g'\frac{\partial\overline{\theta}_w}{\partial s} < 1$ to ensure that s^* is unique.

Comparative static exercises using Eqn (19.39) show how stigma that arises from statistical discrimination responds to changes in various exogenous variables. For example, Besley and Coate show that $\frac{\partial s^*}{\partial b}$ can be positive or negative depending on the value of $\hat{\theta}$. The algebra is tedious and will be left to the interested reader.

A comparative static exercise that relates directly to Besley and Coate's analysis of workfare is the response of stigma to a change in the proportion of the deserving poor, γ. Differentiating Eqn (19.39) with respect to γ and rearranging the terms yields:

$$\frac{\partial s^*}{\partial\gamma} = \frac{g'\frac{\partial\overline{\theta}_w}{\partial\gamma}}{\left(1 - g'\frac{\partial\overline{\theta}_w}{\partial s}\right)} \tag{19.41}$$

The denominator is positive by assumption. Also,

$$\frac{\partial\overline{\theta}_w}{\partial\gamma} = \left(\overline{\theta} - \overline{\theta}_u\right)\frac{\partial\pi}{\partial\gamma} \tag{19.42}$$

which is negative since π is increasing in γ. Therefore, stigma decreases as the proportion of the deserving poor on public assistance increases, the expected result.

The government can directly affect γ in two ways. One is to engage in monitoring the welfare rolls in an effort to detect the undeserving poor. Suppose monitoring can detect a proportion λ of the undeserving poor. Then

$$M(\lambda) = \mathrm{n}\left[\gamma + (1-\lambda)(1-\gamma)\left(1 - \hat{\theta}\right)\right] \tag{19.43}$$

Also,

$$\frac{\partial s^*}{\partial\lambda} = \frac{g'\frac{\partial\overline{\theta}_w}{\partial\lambda}}{\left(1 - g'\frac{\partial\overline{\theta}_w}{\partial s}\right)} \tag{19.44}$$

which has the same negative sign as $\frac{\partial s^*}{\partial\gamma}$. The government can reduce stigma through monitoring.

Workfare is another possibility. If it were designed effectively as described in the previous section, then it would remove all the undeserving poor from the welfare rolls. With $\overline{\theta}_w = \overline{\theta}, \overline{g}(0) = 0$ and stigma disappears.

These examples introduce some doubt about the wisdom of reducing or eliminating welfare stigma when compared with the alternatives. Monitoring is costly and can generate its own form of psychic costs to the poor. Workfare forces the deserving poor to work, which may make them much worse off (recall that the government cannot distinguish the deserving from the undeserving poor without monitoring). In the United States at least, the deserving poor are deserving precisely because they are not expected to work. Stigma is undesirable because it lowers the utility of all the poor, but it has the beneficial effect of keeping some of the undeserving poor off the welfare rolls. Therefore, the nonpoor may well prefer stigma to either monitoring or workfare as a means of reducing the number of undeserving poor on welfare, at least if stigma arises from this kind of statistical discrimination.

Note, finally, that neither the preferences of the nonpoor nor the total number of poor affect stigma when it results from statistical discrimination. The preferences of the nonpoor only come into play in the Besley–Coate model if taxpayer resentment is the source of the stigma.

A Political Note

Assar Lindbeck et al. published a variation of the Besley–Coate model in 1999, in which their main contribution was to add a political dimension (Lindbeck et al., 1999). Their model assumes that individuals have identical preferences but a continuum of skills (wages). There is no sharp distinction between the poor and nonpoor as in the Besley–Coate model. The stigma in their model depends on the number of people who choose to accept transfers rather than work. The polity is a direct democracy in which the transfer-tax policy is decided by a simple majority of voters. Lindbeck et al. assume that everyone votes, so that the voter with the median preferences is decisive.[10] With identical preferences and a continuum of skills, the person with the median skill level is the decisive voter.

Their model yields a number of interesting results, particularly regarding the possibility of multiple equilibria.

10. Chapter 28 has a discussion of the median voter political model.

But we mention it primarily to highlight the additional difficulties economists face as they try to bring political considerations into their models. In the baseline model of Lindbeck et al., with stigma but no altruism, the only possible equilibria under simple majority voting are as follows:

1. Zero taxes and transfers if the majority of voters choose to work and pay taxes.
2. A tax-transfer equilibrium if the majority of voters choose not to work.

Either outcome is far from reality in any of the highly developed market economies. Also, the poor do not vote in anywhere near the same proportion as the nonpoor in the United States, so that a direct democracy with full voting would not seem to be the appropriate political model for determining transfer payments in the United States. Unfortunately, no other obvious alternative comes to mind. Lindbeck et al. do obtain more realistic possibilities with extensions of their baseline model. But the point remains that the assumed political environment can have a dramatic impact on the implications of any economic model, which makes the uncertainties surrounding the appropriate political model all the more troublesome for normative public sector theory.

REFERENCES

Bearse, P., Glomm, G., Janeba, E., March 2000. Why poor countries rely mostly on redistribution in-kind. Journal of Public Economics 75 (3), 463–481.

Besley, T., Coate, S., April 1995. The design of income maintenance programs. Review of Economic Studies 62 (2), 187–222.

Besley, T., Coate, S., July 1992a. Understanding welfare stigma: taxpayer resentment and statistical discrimination. Journal of Public Economics 48 (2), 165–183.

Besley, T., Coate, S., March 1992b. Workfare versus welfare: incentive arguments for work requirements in poverty-alleviation programs. American Economic Review 82 (1), 249–261.

Blackorby, C., Donaldson, D., September 1988. Cash versus kind, self-selection, and efficient transfers. American Economic Review 789 (4), 691–700.

Browning, E., March 1995. Effects of the EITC on Income and Welfare. National Tax Journal 48 (1), 23–43.

Browning, E., Browning, J., 1983. Public Finance and the Price System, second ed. Macmillan, New York.

Bruce, N., Waldman, N., December 1991. Transfers in kind: why they can Be efficient and Nonpaternalistic. American Economic Review 81 (5), 1345–1351.

Buchanan, J., 1975. The Samaritan's dilemma. In: Phelps, E. (Ed.), Altruism, Morality and Economic Theory. Russell Sage Foundation, New York.

Falk, G., October 17, 2013. The Temporary Assistance for Needy Families (TANF) Block Grant: Response to Frequently Asked Questions. Congressional Research Service, Table B.5. www.fas.org/sgp/crs/misc/RL32760.pdf.

Lindbeck, A., Nyberg, S., Weibull, J., February 1999. Social norms and economic incentives in the welfare state. Quarterly Journal of Economics 114 (1), 1–35.

Moffitt, R., 1983. An economic model of welfare stigma. American Economic Review 73 (5), 1023–1035.

Rowe, N., Wooley, F., May 1999. The efficiency case for universality. Canadian Journal of Economics 32 (3), 613–629.

Schloz, J., March 1994. The Earned Income Tax Credit: Participation, Compliance, and Anti-poverty effectiveness. National Tax Journal 47 (1), 63–87.

Chapter 20

Social Insurance: Medical Care

Chapter Outline

The industrialized market economies provide substantial amounts of public or social insurance, primarily medical insurance, retirement annuities, and unemployment insurance. The United States is certainly no exception. In Fiscal Year 2010, expenditures for Medicare and Medicaid, the public insurance programs for the elderly and the poor, exceeded $800 billion, Social Security pension benefits to retirees exceeded $600 billion, and unemployment insurance was just under $160 billion. Together, these programs accounted for approximately 30% of total federal, state, and local government expenditures. At the same time, all the industrialized market economies have well developed private insurance markets, for life insurance, automobile insurance, property insurance, and even medical insurance and private annuities, despite the presence of large public programs in these areas.

This chapter considers the factors that can make private insurance markets inefficient and even threaten their very existence, and the factors that lead to a demand for social insurance. It also discusses the optimal policy responses to each factor. The topic is inherently part of second-best expenditure theory since one of the more important factors is private or asymmetric information, information that the insured have about themselves that insurers cannot know, or at least do not know well enough to confidently offer insurance. The focus of this chapter is on the consequences of private information, using medical insurance as an example. Chapter 21 discusses the public provision of retirement annuities, centered on the U.S. Social Security System. Since the market failures associated with private information are inherently efficiency issues, we will ignore distributional concerns in the next two chapters and assume that the government is redistributing income lump sum to equalize social marginal utilities of income.

THE DEMAND FOR INSURANCE

Individuals demand insurance because they are risk averse and want protection against the various misfortunes that life can present to them. They act as if they have diminishing marginal utility of income, as depicted in Fig. 20.1.

Suppose a misfortune can occur with some positive probability. Income is Y_L if the misfortune occurs, and Y_H if it does not. A transfer of income, d, across the two states of nature always increases utility under diminishing marginal utility of income: $[U(Y_L + d) - U(Y_L)] > [U(Y_H) - U(Y_H - d)]$, as illustrated in the figure. Smoothing income in the face of uncertain misfortune is precisely what insurance allows people to do. They sacrifice income if misfortune does not occur by paying a premium to an insurance company in return for receiving a payout from the company if the misfortune does occur. This transfer of income across states of nature raises expected utility in an uncertain environment. The income

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00020-0

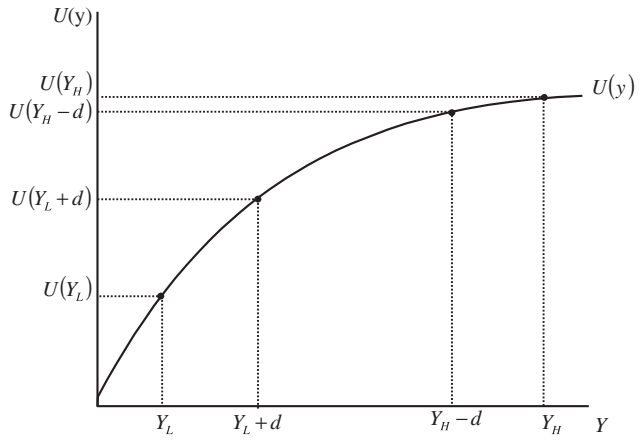

FIGURE 20.1

transfer is commonly referred to as consumption smoothing or, equivalently, risk reduction in the insurance literature.

The following simple example illustrates some fundamental principles of insurance that we will need in order to understand how insurance markets become vulnerable to private information. Suppose all individuals are identical in the following three respects: (1) They are equally risk averse, with utility functions in terms of income as pictured in Fig. 20.1; (2) They all face the same probability, π, of becoming ill, and therefore $(1 - \pi)$ is the probability of remaining healthy; and (3) If healthy, they have income Y. If they become ill, they lose L dollars of income and have income $(Y - L)$.

WITHOUT INSURANCE

Figure 20.2 illustrates the expected loss, the expected income, and the expected utility of an individual without insurance.

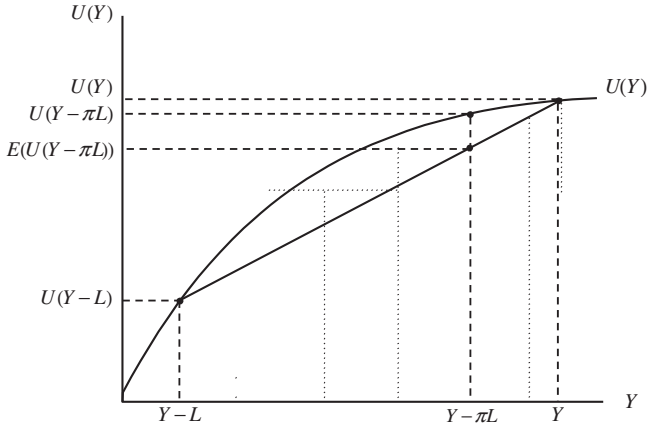

FIGURE 20.2

a. *The expected loss*: $E(\text{loss}) = \pi L$, the probability of becoming ill times the amount of the loss, if ill.

b. *Expected income*: $E(Y) = (1 - \pi)Y + \pi(Y - L)$
$= Y - \pi L$

c. *Expected utility*: $E(U) = (1 - \pi)U(Y) + \pi U(Y - L)$.

The expected utility lies on a straight line segment between $U(Y)$ and $U(Y - L)$, with the actual expected utility depending on the value of π. The end points occur with probabilities $\pi = 0$ $[U(Y)]$ and $\pi = 1$ $[U(Y - L)]$. The actual expected utility lies directly above the expected income, as illustrated in the figure.

WITH INSURANCE

The purchase of insurance allows individuals to increase their expected utility by smoothing their income (consumption) across the two states of nature, ill and healthy.

Actuarially fair, full insurance—The pareto-optimal solution for each individual is to be able to purchase actuarially fair, full insurance. *Actuarially fair insurance* means the individual pays a premium, P, equal to the expected payout under the policy. If X is the payout, the expected payout is πX. Thus the insurance policy is actuarially fair if $P = \pi X$. Full insurance means that the payout equals the loss, $X = L$. Therefore, under actuarially fair, full insurance $P = \pi L$, the premium equals the expected loss.

The Supply Side

A brief detour to the supply side of the market is in order at this point. For insurance companies (insurers) to be able to offer medical insurance, two conditions must hold. First, the insurers must know the risk, π, of the insured, i.e., the probability of becoming ill. Second, the otherwise identical individuals in our example must differ in this respect: The probability that any one individual becomes ill must be independent of the probability of any other individual becoming ill. If this is true, and the insurance companies can each insure a large number of individuals, then the law of large numbers implies that actual payouts will equal the expected payouts across the number of individuals insured. This is the so-called pooling effect of insuring a large number of individuals whose risks are statistically independent. Therefore, an actuarially fair premium equal to the expected payout collected from a large number of individuals allows the firm to break even. This assumes no administrative costs, called loads in the insurance industry, which we will continue to assume throughout this chapter unless loads are specifically considered. Adding loads does not materially affect most of the analysis. If insurance markets are competitive, as they must be for a pareto-optimal outcome, then each company must break even in

equilibrium. The break-even condition in competitive insurance markets is

$$\sum_{i=1}^{N} P_i = \sum_{i=1}^{N} \pi X_i \tag{20.1}$$

where N is large enough for the law of large numbers to apply. With full insurance,

$$\sum_{i=1}^{N} P_i = \sum_{i=1}^{N} \pi L_i \tag{20.2}$$

In our simple model with identical individuals, Eqn (20.2) is $NP = N\pi L$, or $P = \pi L$. Actuarially fair, full insurance is a competitive equilibrium.

The independence assumption does not necessarily apply, however. Millions of people become unemployed when the economy goes into a recession and millions of these same people become employed again when the economy recovers. As a result, the law of large numbers does not apply and the risk of unemployment is not insurable. An insurance company offering unemployment insurance would face much the same risk faced by any one individual, multiplied many times over. Therefore, if people want unemployment insurance, the government has to provide it and assume the risks. The same statistical dependence inhibits the provision of insurance against flooding when riverbanks overflow or levies are breached.

The Pareto-Optimum

If the risks are independent, however, and insurance markets are competitive, then actuarially fair, full insurance is both a competitive equilibrium and pareto-optimal for the individuals. Under actuarially fair, full insurance, each individual receives an income of $U(Y - P) = U(Y - \pi L)$, if healthy with probability $(1 - \pi)$ and $U(Y - L - P + L) = U(Y - P) = U(Y - \pi L)$, if ill with probability π. Since income is the same in both states, income of $Y - \pi L$ is received with certainty and each individual achieves $U = U(Y - \pi L)$, the point on the utility function that lies directly above $Y - \pi L$, the expected income under uncertainty. Actuarially fair, full insurance thus provides the maximum possible benefit of consumption smoothing in an uncertain environment by removing the uncertainty.

This result is an application of the Bernoulli's theorem: A risk-averse individual prefers any income level Y^* with certainty to an uncertain income with an expected value of Y^*.

The value to the insured of actuarially fair, full insurance is the distance between the utility function and the expected utility line above the expected income $Y - \pi L = Y - P$, $U(Y - P) - E[U(Y - P)]$ in Fig. 20.2. To approximate the value, perform a Taylor series expansion of $E(U) = (1 - \pi)U(Y) + \pi U(Y - L)$ around the point $(Y - P)$, income net of the premium.

$$\begin{aligned} E(U) = {} & (1-\pi)\left[U(Y-P) + U'P + 1/2U''P^2\right] \\ & + \pi\left[U(Y-P) - U'(L-P) + 1/2U''(L-P)^2\right] \end{aligned} \tag{20.3}$$

which simplifies to[1]

$$E(U) = U(Y-P) + 1/2U''P(L-P) \tag{20.4}$$

Therefore, $U(Y - P) - E(U(Y - P)) = -1/2U''P(L - P)$. Dividing by U' to express the difference in dollar terms,

$$\frac{U(Y-P) - E(U(Y-P))}{U'} = -\frac{1}{2}\frac{U''}{U'}P(L-P) \tag{20.5}$$

$-\left(\frac{U''}{U'}\right)$ is the coefficient of absolute risk aversion, reflecting the curvature of the utility function. The term $P(L - P)$ reflects the difference in income between being fully insured and uninsured. P is the loss in income when healthy by being insured and $(L - P)$ is the increase in income when ill by being insured. The value of being insured, therefore, depends on the amount of risk aversion and the amount of income (consumption) smoothing resulting from the insurance.

Actuarially fair, partial insurance—Actuarially fair, partial insurance is also valued by risk-averse individuals, but less so than full insurance. To see this, let the payout $X = aL$, where $0 < a < 1$, and the actuarially fair premium $P = \pi aL$. Expected income with partial insurance is

$$\begin{aligned} E(Y) &= (1-\pi)(Y - \pi aL) + \pi(Y - L - \pi aL + aL) \\ &= (1-\pi)Y - (1-\pi)\pi aL + \pi(Y-L) \\ &\quad + \pi(1-\pi)aL \\ &= (1-\pi)Y + \pi(Y-L) = Y - \pi L \end{aligned}$$

The insurance terms cancel and expected income is the same as with no insurance (or with actuarially fair, full insurance with certainty). So long as insurance is actuarially fair, it does not change the expected income:

$$\begin{aligned} E(U) &= (1-\pi)U(Y - \pi aL) + \pi U(Y - L - \pi aL + aL) \\ &= (1-\pi)U(Y - \pi aL) + \pi U(Y - L + (1-\pi)aL) \end{aligned}$$

Refer to Fig. 20.3. The expected utility line retains the same slope as the expected utility line without insurance in Fig. 20.2, but the end points of the expected utility line

1. $E(U) = U(Y - P) + U'((1 - \pi)P - \pi(L - P)) + 1/2U''((1 - \pi)P^2 + \pi(L - P)^2)$. $P = \pi L$, so that the term in parentheses following U' is zero. The term in parentheses following U'' equals $(1 - \pi)P^2 + \pi L^2 - 2\pi LP + \pi P^2$. With $P = \pi L$, this term equals $\pi L^2 - P^2$ or $P(L - P)$.

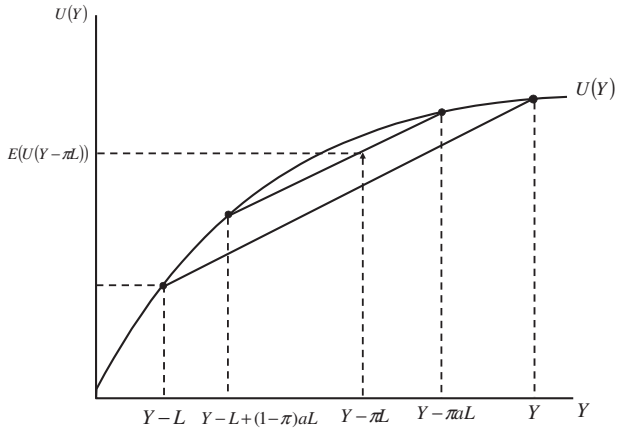

FIGURE 20.3

have moved closer together. This raises expected utility relative to having no insurance as illustrated by the figure, but it falls short of the utility achieved with full insurance.

The risk premium—So far we have considered actuarially fair insurance. Risk-averse individuals are willing to purchase full insurance that is not actuarially fair, however, as illustrated in Fig. 20.4.

They have expected utility of $E(U(Y - \pi L))$ without insurance. Therefore they will pay a premium for full insurance, provided that they have a utility level with certainty that equals or exceeds $E(U(Y - \pi L))$. The point of indifference in the figure is $Y - P - bP = Y - (1 + b)P$. The amount bP is called the *risk premium*, the maximum amount the individual is willing to pay above the actuarially fair premium to obtain full insurance. The insured's willingness to pay more than an actuarially fair premium up to the amount of the risk premium is what allows insurance companies to cover their administrative costs and earn a return to capital equal to opportunity cost of capital when offering full insurance coverage. The break-even condition for the insurers may not include the full risk premium. We will ignore the risk premium and insurance loads for most of the rest of this chapter.

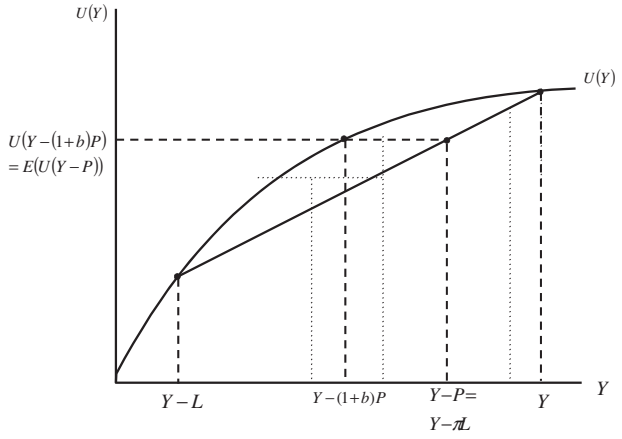

FIGURE 20.4

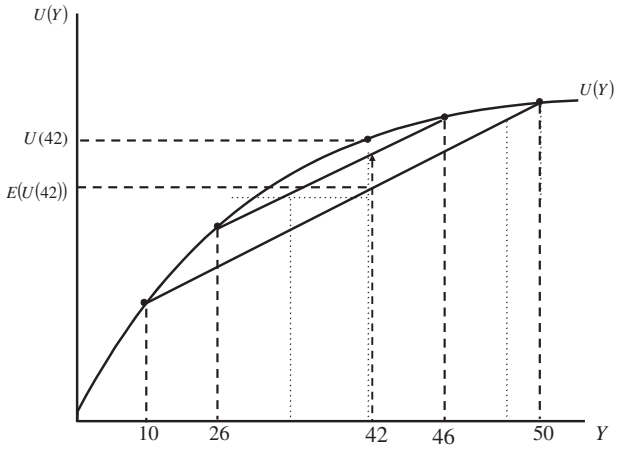

FIGURE 20.5

Numerical example—A numerical example will illustrate these principles (except for the risk premium). Suppose all the individuals have $Y = \$50$ when healthy, and lose \$40 when ill, for an income of $Y - L = \$10$. The probability of becoming ill, π, is 0.2.

Without insurance:

$$E(\text{Loss}) = \pi L = 0.2(\$40) = \$8$$

$$E(Y) = 0.8(\$50) + 0.2(\$10) = \$42$$

$$= Y - E(\text{Loss})$$

$$E(U) = 0.8U(\$50) + 0.2U(\$10),$$

Refer to Fig. 20.5. $E(U)$ is above $E(Y) = \$42$ on the lower expected utility line, whose end points are on $U(Y)$ at \$50 and \$10.

With actuarially fair, full insurance, the individual receives $Y = \$42$ with certainty. The premium is \$8, equal to the expected loss. Therefore, with $\pi = 0.8$, the individual has an income of \$42 (\$50 − \$8) and also with $\pi = 0.2$, the individual has an income of \$42 (\$50 − \$40 − \$8 + \$40). Income is \$42 in either state of nature, and utility is U(\$42), with certainty.

Suppose, instead the individual purchases actuarially fair, partial insurance that covers 50% of the loss. The premium is 0.2(0.5)(\$40) = \$4 and the payout is \$20, if ill.

$E(Y) = 0.8(\$50 - \$4) + 0.2(\$50 - \$40 - \$4 + 20) = 0.8(\$46) + 0.2(\$26) = \$36.8 + \$5.2 = \$42$, the same as with no insurance (or full insurance). $E(U)$ is above \$42 on the expected utility line whose end points are \$46 and \$26, a higher $E(U)$ than without insurance but less than U(\$42) attainable with full insurance.

PRIVATE OR ASYMMETRIC INFORMATION

Perfect information is a common assumption in market analysis, meaning that the buyers and sellers have sufficient

information about the goods and services to confidently engage in exchange. Quite often, however, the information between buyers and sellers is asymmetric. One side has private information that the other side cannot know, or cannot know without engaging in costly monitoring or testing. Producers typically know more about their goods and services than do individuals, and this is certainly true in the market for medical care. Physicians have an enormous informational advantage over their patients regarding their illnesses and the efficacy of different medical procedures and drugs. Because of the nature of the service, however, this does not usually prevent individuals from seeking out physicians when they become ill or injured. But the informational advantage rests with the individuals in the market for medical insurance, and in many other insurance markets as well. Individuals often have private information about the risks that they are asking insurance companies to insure against. In the context of medical insurance, the insurers can find out information about individuals' medical histories, but they cannot easily determine whether they lead relatively healthy or unhealthy lifestyles.

If the information on risks is poor enough, private insurers become vulnerable to two problems, moral hazard and adverse selection. The classic definition of *moral hazard* is that the insured can take actions, unbeknownst to the insurers, that change the probability of the risks being insured against. Insurers are understandably reluctant to offer insurance if they are vulnerable to this kind of behavior. *Adverse selection* arises if different individuals present different risks to the insurers, but the insurers cannot differentiate the insured according to risk. As a result, they must charge a single premium to all the insured. The problem here is that low-risk individuals never want to be pooled with high-risk individuals, with the result that the low-risk individuals might not be able to obtain the insurance coverage they want and are willing to pay for. Even worse, the insurance market can completely unravel and have no equilibrium, such that no one can obtain insurance. We will consider each of these problems in turn.

MORAL HAZARD

A simple variation of our original model, developed by Mark Pauly, illustrates the classic case of moral hazard.[2] Assume, as above, that all individuals are identical and face a probability π of suffering an illness that lowers their income. The difference here is that individuals can take a preventive action, Z, that lowers the probability of the illness occurring $d\pi/dZ < 0$. The price of Z is one. Think of the insured incurring additional expenses to eat a healthier diet or to join a health club to get more exercise. The insurers cannot observe Z. All they can do is set the premium, P, and the payout, X, on their policies knowing that the individuals will adjust Z in response to both.

The maximization of expected utility is this model occurs in two stages. First, the individuals determine the optimal amount of Z, taking the premium, P, and the payout, X, as given. Then the insurers determine the amount of P and X that maximizes expected utility, given the individuals' reaction function of Z to P and X.

Consider, first, the decision on Z:

$$\underset{(Z)}{\text{Max}}\, E(U) = (1-\pi)U(Y-P-Z) + \pi U(Y-L-P-Z+X)$$

The FOC are

$$\frac{\partial E(U)}{\partial Z} = -(1-\pi)U'(H) - \pi U'(I) - U(H)\frac{\partial \pi}{\partial Z} + U(I)\frac{\partial \pi}{\partial Z} = 0 \tag{20.6}$$

where $U(H)$ and $U(I)$ are utilities when healthy and ill. Rearranging terms,

$$(1-\pi)U'(H) + \pi U'(I) = (U(I)-U(H))\frac{\partial \pi}{\partial Z} \tag{20.7}$$

The left-hand side is the marginal cost of Z expressed in terms of the loss in expected marginal utilities. The right-hand side is the marginal benefit of Z, the marginal increase in expected utility by reducing the probability of loss as utility is transferred from the ill state to the healthy state.

The insurers then maximize $E(U)$ with respect to the payout, X, under the assumption that both P and Z are functions of X. Furthermore, a natural assumption is that $\frac{\partial Z}{\partial X} < 0$: Individuals reduce preventive activity as the insurance payout increases. Indeed, as Pauly notes, the basis of moral hazard is that an individual assumes that changes in Z do not affect the premium because he or she is just one of a very large number of people being insured by any one insurance company, whereas the expected payout does depend on Z:[3]

$$\underset{(X)}{\text{Max}}\, E(U) = (1-\pi)U(Y-P-Z) + \pi U(Y-L-P-Z+X)$$

The FOC are

$$\frac{\partial E(U)}{\partial X} = \pi U'(I) - \frac{\partial P}{\partial X}[(1-\pi)U'(H) + \pi U'(I)] - \frac{\partial Z}{\partial X}[(1-\pi)U'(H) + \pi U'(I)] + \frac{\partial \pi}{\partial Z}\frac{\partial Z}{\partial X}[U(I)-U(H)] = 0 \tag{20.8}$$

2. Pauly (1974). The analysis in this section closely follows the analysis of moral hazard on pp. 44–54.

3. *Ibid.*, p. 48.

From (20.7), the last two terms sum to zero. Therefore,

$$\frac{\partial P}{\partial X}[(1-\pi)U'(H)+\pi U'(I)] = \pi U'(I) \tag{20.9}$$

or

$$\frac{\partial P}{\partial X} = \frac{\pi U'(I)}{[(1-\pi)U'(H)+\pi U'(I)]} \tag{20.10}$$

Equation (20.10) gives the optimal marginal pricing schedule for the insurance policy, given the reaction function of the individuals' preventive activity. It is a second-best optimum because the insurers cannot observe Z.

To see that it is second best, assume that the insurers could observe Z so that the premium can be made to depend on Z. Knowing that the premium depends on Z, the individual now solves the following problem in the first stage:

$$\begin{aligned}\text{Max } E(U) = {} & (1-\pi)U(Y-P(Z)-Z) \\ & +\pi U(Y-L-P(Z)-Z+X)\,(Z)\end{aligned}$$

The FOC are

$$\begin{aligned}&-(1-\pi)U'(H)\left[1+\frac{\partial P}{\partial Z}\right]-\pi U'(I)\left[1+\frac{\partial P}{\partial Z}\right] \\ &+[U(I)-U(H)]\frac{\partial \pi}{\partial Z} = 0\end{aligned} \tag{20.11}$$

Equation (20.11) is zero if

a. $U(I) = U(H)$, which is true only with full insurance and

b. $1+\frac{\partial P}{\partial Z} = 0$. 1 is the marginal cost of Z. If the insurance is actuarially fair, then $P = \pi X$, which equals πL under full insurance. Therefore, $\frac{\partial P}{\partial Z} = \frac{\partial \pi}{\partial Z}L$, the marginal reduction in the expected loss, which is the marginal benefit of Z. The individuals purchase Z such that the marginal benefit of Z equals its marginal cost. In effect, being able to observe Z is equivalent to being able to observe π.

The first-best solution obtains, with the pareto-optimal amount of insurance for the individuals and each individual engaging in the optimal amount of preventive activity. This is a particular instance of a quite general result in the insurance literature: If insurers can observe preventive activities taken before the state of nature is revealed that affect the state of nature (or the payouts), then they can offer first-best insurance policies and preserve first-best incentives for the preventive activities. This is true of a wide variety of insurance models.

At the second-best optimum given by Eqn (20.10), the individuals engage in too little Z, such that $MB_Z > MC_Z$, but the difference is compensated for by an increase in X occasioned by purchasing too little Z (with $\frac{\partial Z}{\partial X} < 0$). From the insurers' point of view, they would like the marginal prices given by $\frac{\partial P}{\partial X}$ to equal the marginal change in actual claims, X, by the insured. Because they cannot observe Z, however, the best they can do is to set the marginal price schedule equal to the change in the expected claims or payout. To see this, reintroduce the assumption that the insurance is actuarially fair, $P = \pi X$, and substitute πX for P in the maximization with respect to X above. The result would be

$$\left[\pi+\frac{\partial \pi}{\partial X}X\right] = \frac{\pi U'(I)}{[(1-\pi)U'(H)+\pi U'(I)]} = \frac{\partial P}{\partial X} \tag{20.12}$$

$\left[\pi+\frac{\partial \pi}{\partial X}X\right]$ is the change in the expected payout with respect to X.

THE COMPETITIVE OUTCOME

The second-best price schedule will almost certainly not arise in a competitive insurance market, however, for a number of reasons. First, each insurer would need to know the entire amount of insurance that any individual buys to implement the second-best price schedule, but insurers are likely to know only how much insurance each of their policyholders buys from them. Second, if an insurer tried to increase premiums with the amount of insurance purchased, the insured would buy only the first unit of insurance from that firm, and buy additional units from other firms. Finally, competitive firms are subject to the break-even condition, which, for N consumers and actuarially fair insurance, is $\sum_{i=1}^{N} P_i = \sum_{i=1}^{N} \pi X_i$. Assuming all individuals have identical risk, π, the price per unit of insurance, p, equals $\frac{\sum_{i=1}^{N} \pi X_i}{\sum_{i=1}^{N} X_i} = \pi$, which would be identical for all insurers since they are price takers. With $\frac{\partial P}{\partial X} = P = \pi$, the marginal price is less than the second-best optimal level of $(\pi+\frac{\partial \pi}{\partial X}X)$ with $\frac{\partial \pi}{\partial X} > 0$, and the individuals insured buy too much insurance. Indeed, they buy much too much. With $p = \pi$, Eqn (20.10) becomes

$$p = \frac{pU'(I)}{[(1-p)U'(H)+pU'(I)]} \tag{20.13}$$

or

$$(1-p)U'(H) = (1-p)U'(I) \tag{20.14}$$

Equation (20.14) implies $U(H) = U(I)$, or full insurance. Moreover $Z = 0$. This follows from Eqn (20.7). With full insurance, the RHS is zero and the LHS is $U' > 0$.

Spending on Z would sacrifice utility without any benefit. Intuitively, since individuals assume they cannot affect the price (premium), and $p = \pi$, they have no incentive to lower π. Competitive insurance outcomes can be highly inefficient in the presence of moral hazard.

THE PUBLIC POLICY RESPONSE

The government presumably is no more able to observe the individuals' preventive actions than are the private insurers. Therefore, the best it can do to maximize individual utility is the second-best outcome given by Eqn (20.10). As noted above, implementing the second-best marginal price schedule $\frac{\partial P}{\partial X}$ requires knowing the total amount of insurance purchased by each individual. Therefore, the government must require individuals to report their total insurance purchases. Assume it does so and the requirement is enforceable. Then it can provide insurance in accordance with Eqn (20.10). But government provision has no advantage over the private insurers in this regard since each firm can also implement the second-best marginal price schedule if it knows the total purchases of the insured. The firm would set the premium for each increment of X equal to the incremental expected payouts at each X, based on its knowledge of $\pi(X)$. With all firms having the same information, they all implement the same price schedule. The individuals, in turn, would base their purchases on a schedule that is the margin of the optimal $\frac{\partial P}{\partial X}$ supply schedule, and the market would achieve a break-even equilibrium at the second-best optimum.

EX POST MORAL HAZARD[4]

The classic definition of moral hazard in which individuals can influence the probability of a payout is often referred to as *ex ante* moral hazard. The definition of moral hazard has also been expanded to include the consequences of overusing medical care that has been effectively priced too low because the medical care is insured. This is referred to as *ex post* moral hazard. Of the two, ex post moral hazard is undoubtedly the more important in the market for medical care in terms of the amount of inefficiency that results. Engaging in unhealthy lifestyles has its own direct utility costs that many people choose to avoid (these costs could be captured by introducing health into the utility function). But ex post moral hazard can generate substantial inefficiencies, as illustrated in Fig. 20.6.

The horizontal axis is units of medical care and the vertical axis is dollars. Suppose the demand curve for medical care if an individual becomes ill is the solid line labeled D. The marginal cost of providing medical care is MC, the supply curve, assumed to be constant. With no insurance, the individual buys 100 units of medical care, at the intersection of D and MC. The probability of becoming ill is $^1/_2$. Therefore, the expected loss of becoming ill is $^1/_2$ (100)MC = 50 MC.

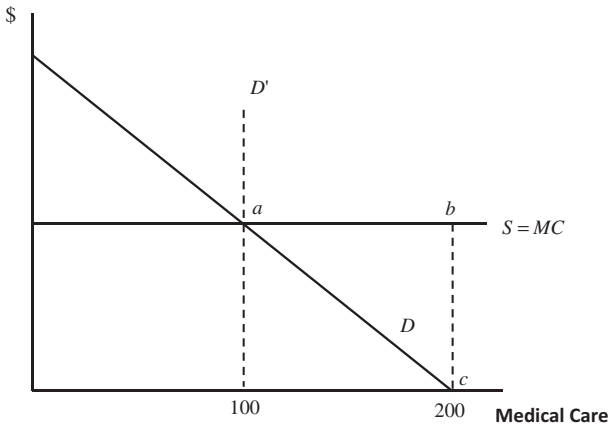

FIGURE 20.6

Suppose the demand curve for medical care were perfectly inelastic at 100 units of care, the dotted line D′. Then, if the individual receives a full insurance policy, the premium would be 50MC, equal to the expected payout under the policy, and the individual's net income would be $Y - 50$MC whether healthy or ill. If individuals are risk averse, they would rather pay 50MC with certainty than face an expected loss of 50MC without insurance, as we saw above (a restatement of the Bernoulli's theorem above in terms of the certain premium and the expected loss).

If the demand for medical care is at all elastic, however, as with demand curve D above, the advantage of full insurance is less clear. The problem is that the marginal cost to any one individual of purchasing medical care is zero under full insurance. Therefore, the insured overuses the insurance, purchasing 200 units of medical care in the figure at a perceived price of zero. But the cost of 200 units is 200MC, and the insurers have to cover those costs by charging higher premiums. The premium rises to 100MC [=1/2(200)MC]. Individuals may well prefer to face an expected loss of 50MC then sacrifice 100MC in each state with certainty.

Another way to see the inefficiency from ex post moral hazard is to assume that there are no income effects, so that D is both the actual and compensated demand curve. The loss to the individual of overusing the insurance is the area abc of negative consumer surplus, equal to the cost of the insurance from units 100 to 200 less the value to the insured of receiving an additional 100 units of medical care, the area under the demand curve from 100 to 200 units. The

4. The analysis of ex post moral hazard is a variation of the analysis in Pauly (1968).

loss may well be greater than the benefit from consumption smoothing that full insurance permits. Indeed, empirical estimates suggest that the loss is generally much larger than the gains from consumption smoothing for any reasonable degree of risk aversion.[5]

The insured are caught in the Prisoner's Dilemma. The pareto-optimal outcome is for each to have full insurance and use 100 units of medical care if illness occurs. But from an individual perspective, the best move appears to be to buy 200 units at the zero price and spread the costs to everyone else. Individuals correctly assume that their actions, by themselves, have no appreciable effect on costs and premiums because they are each one among many. That is, they measure the direct gain myopically as the area under the demand curve from units 100 to 200. Since everyone thinks that way, however, they all purchase 200 units of medical care and raise the costs and premiums, generating inefficiency to the point that the insurance may not be worthwhile.

DEDUCTIBLES AND CO-PAYMENTS

A common strategy among private and government insurers to reduce the extent of the ex post moral hazard is to require deductibles and co-payments. A deductible of $\$X$ means that the insured pays the first $\$X$ of the medical costs. By itself, a deductible leads to an uncomfortable all-or-none outcome, at least in the model we have been using. It either has no effect on the purchase of insurance or removes the demand for insurance entirely as Fig. 20.7 illustrates.

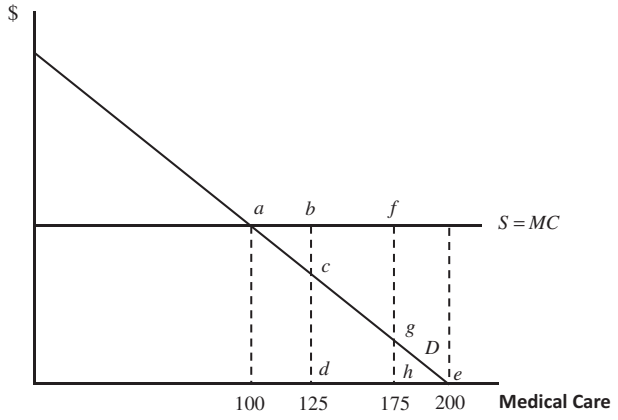

FIGURE 20.7

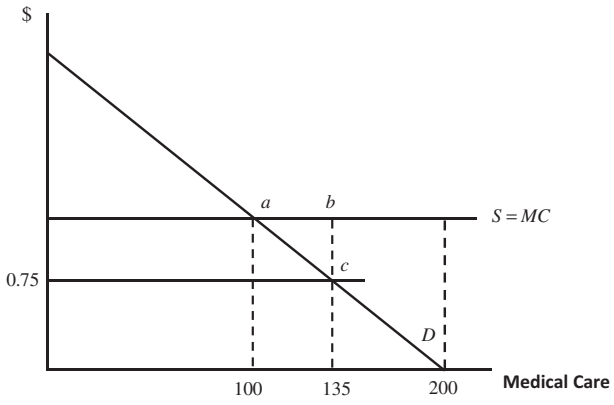

FIGURE 20.8

Suppose the insured receive full insurance subject to the deductible that they pay for the first 125 units of medical care. The deductible raises the costs to the insured by area abc, equal to the cost of paying for units 100 through 125 less the area under the demand curve between 100 and 125. The gain from being able to use units 125–200 for free is viewed, myopically, as area cde, the area under D from 125 to 200. If abc $<$ cde, as in the figure, then the deductible has no effect on the insured's behavior. They continue to purchase 200 units. A large enough deductible, however, say the first 175 units, reverses the inequality. Area afg, the extra costs to the insured, exceeds area egh, the extra benefits of being able to use units 175–200 for free, and the individual buys no insurance.

Co-payments are potentially more useful. Refer to Fig. 20.8.

Suppose the insured have to pay 75% of the costs of the medical care, the horizontal line labeled 0.75 in the figure. The quantity of medical care demanded decreases with that co-payment, to 135 units in the figure. The net loss to the insured is area abc, which may be less than the value of consumption smoothing from having insurance coverage for 135 units of medical care. The co-payments, by reducing the moral hazard incentive to overuse insurance, could make insurance worthwhile when it is not under full insurance with a perceived price of zero. The optimal co-payment rate is that which just balances on the margin the benefits from reducing the inefficiency from over-consumption and the costs of reduced consumption smoothing from having to pay more for medical care.

THE VALUE OF ACCESS

John Nyman argues persuasively that the standard analysis in the literature of the gains and losses from insurance resulting from ex post moral hazard, the analysis given above, is inaccurate in a number of respects and may reach the wrong conclusions. First, he notes that income effects are likely to be important in the demand for medical care,

5. John Nyman lists a number of papers that estimate the relative value of the consumption smoothing benefits and ex post moral hazard costs for different insurance policies on p. 142 in Nyman (1999b). He cites two of the results from these papers, one which found that the ex post moral hazard costs were 10 times the consumption smoothing benefits, the other two times the consumption smoothing benefits.

such that the appropriate compensated demand curve for measuring gains and losses lies inside the actual demand curve. This matters, because a portion of the actual demand for medical care under full insurance results from a transfer of income from the healthy to the ill, a transfer that supports the demand for medical care as well as the demand for other goods and services when ill. The higher the transfer the lower is the probability of an illness. Properly measured, therefore, the inefficiency from full insurance with a perceived price of zero would be smaller than measured above because not as much medical care is demanded along the properly measured demand curve. Only the substitution effect (the price effect) matters in measuring the inefficiency of overusing insurance, not the income effect.

Second, the standard analysis misses a very important gain besides consumption smoothing, that insurance gives many people access to medical care they otherwise could not afford. Nyman gives the example of a kidney transplant, which at the time he wrote cost $300,000, well beyond the reach of most individuals. Since the probability of needing a transplant is only 1.33 cases per 100,000 people, the expected payout is only $4 per year $\left[4 = \$300,000\left(\frac{1.33}{100,000}\right)\right]$, the value of the premium under actuarially fair insurance. He claims that the benefits from the access to very expensive medical care that insurance permits are far greater than the benefits of consumption smoothing for reasonable values of risk aversion, perhaps three times as much as a lower bound. Taking the gains from access into account might well make full insurance at a price of zero beneficial even with the moral hazard. If so, then policies such as deductibles and co-payments are welfare reducing, if anything, especially if they are targeted to the most expensive procedures for which the gains from access are the largest.

His final point is that the costs of offering full insurance given moral hazard, even if greater than the benefits, may still be the least cost way of providing transfers to the ill. Lump-sum transfers to those who are ill may appear to be more cost effective in theory than full insurance, but they are more easily susceptible to fraud. In summary, Nyman cautions against dismissing full insurance because of ex post moral hazard.[6]

ADVERSE SELECTION

Adverse selection arises when individuals differ in the risk they present to the insurers—their probabilities, π_i, of becoming ill—and the risks are private information to the insured. Therefore the insurers have to offer a single premium to all the insured, which leads to a number of difficulties. We will assume that moral hazard is not an issue in this section to focus on the consequences of adverse selection.

EX POST VERSUS EX ANTE EFFICIENCY

As it happens, an efficiency issue arises when individuals have different risks even when their risks are known to the insurers. To see the problem, begin as above with a single individual having a probability π of becoming ill, but adopt a slightly different framework. Refer to Fig. 20.9.

The horizontal axis is income if healthy, Y_H and the vertical axis is income if ill, Y_I. The individual has income Y if healthy and $(Y - L)$ if ill, where L is the loss, point A in the figure.

A line from any point such as A below the 45°-line $(Y_H > Y_I)$ with a slope of $\frac{(1-\pi)}{\pi}$ is called the fair-odds line because it represents actuarially fair insurance. Under actuarially fair insurance, $P = \pi X$, where P is the premium and X is the payout. Subtract πP from both sides: $(1 - \pi)P = \pi(X - P)$. Therefore $\frac{(1-\pi)}{\pi} = \frac{(X-P)}{P} = -\frac{dY_I}{dY_H}$, the ratio of the net payout to the premium under actuarially fair insurance.

The individual's expected utility, $E(U) = (1 - \pi)U(Y_H) + \pi U(I)$, generates a set of indifference curves defined by $(1 - \pi)U(Y_H) + \pi U(Y_I) = k$. Consider the slope of an indifference curve:

$$(1 - \pi)U'(Y_H)\mathrm{d}Y_H + \pi U'(Y_I)\mathrm{d}Y_I = 0 \qquad (20.15)$$

$$\frac{-\mathrm{d}Y_I}{\mathrm{d}Y_{H|U=\overline{U}}} = \frac{(1-\pi)U'_H}{\pi U'_I} \qquad (20.16)$$

A fair-odds line that goes to the 45°-line represents actuarially fair, full insurance, such that $Y_H = Y_I$. With equal incomes in both states, $U'_H = U'_I$, the slope of an indifference curve is $\frac{(1-\pi)}{\pi}$ on the 45°-line, the same as the

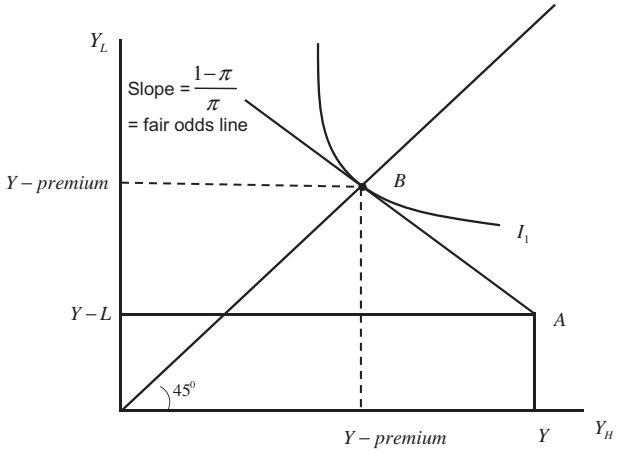

FIGURE 20.9

6. Nyman (1999a), and Nyman (1999b). The kidney transplant example is on pp. 144–145. The estimate of the value of access relative to the value of consumption smoothing is on p. 149.

fair-odds line. Therefore, point B is an equilibrium with actuarially fair, full insurance starting at point A.

Now suppose that there are two groups of individuals identical within each group, one a low-risk group with probability π_L of becoming ill and the other a high-risk group with the probability π_H of becoming ill, $\pi_H > \pi_L$. Otherwise individuals in the two groups have the same incomes and suffer the same loss L if ill so that they are all at point A without insurance. They also have the same expected utility functions. If the insurers can distinguish the individuals by risk, they offer actuarially fair, full insurance to each. Refer to Fig. 20.10. The high-risk individuals end up at point C along the fair-odds line AC with slope $\frac{(1-\pi_H)}{\pi_H}$, paying an actuarially fair premium of $\pi_H L$. The low-risk individuals end up at point D along the fair-odds line AD with slope $\frac{(1-\pi_L)}{\pi_L}$, paying a premium of $\pi_L L$. The insurance policies are pareto-optimal for all individuals ex post, given their risks.

A problem, though, is that high- and low-risk individuals end up paying different premiums for their insurance, whereas ex ante, before the risks are revealed by nature, the optimum would be for each individual to pay the same amount. Different people should not have to pay different amounts for health insurance over their lifetimes just because of the luck of their health draws from nature (recall the assumption in this section of no ex ante moral hazard). There is a missing market that prevents the equal-payments outcome, however, a market for premium insurance. Ideally, all people would be able to pay the same actuarially fair premium for insurance that would cover increases in premiums as they experience bad draws of health. That is, the ex ante optimum is a policy with a single premium that pools everyone's lifetime risks of illness together. The insurance might have to be offered to parents for their children before the children are born. Insurers have not been willing to offer such insurance, however, probably because it is so difficult to determine what the appropriate premium should be. Also, even if they could predict how future illnesses would affect health care costs and thus premiums, the projected increase in costs would affect all the insured and thus not be independent events. Premiums might not be insurable. Consequently, premiums are adjusted after the health draws from nature occur, and the best feasible outcome becomes the ex post optimality pictured in Fig. 20.10.

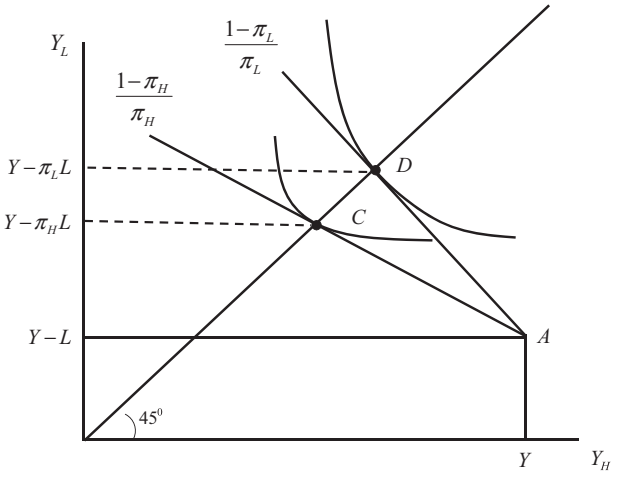

FIGURE 20.10

THE NATURE OF ADVERSE SELECTION

The inability of insurers to distinguish individuals by risk prevents the market from reaching even the ex post optimum. Indeed, at its worst, private information about risks can cause private insurance markets to completely unravel, with no one receiving insurance even though everyone wants insurance and is willing to pay the insurer for the costs they impose on the firm.

Liran Einav and Amy Finkelstein (E/F) developed a simple model of an insurance market to illustrate a fundamental pricing problem under adverse selection that can easily lead to an inefficient allocation of insurance.[7] Assume that T individuals lie on a continuum of probabilities of becoming ill, from highest to lowest risk, and that the risks are unknown to the insurers. Insurers offer only one kind of insurance policy, for which they have to charge the same price to everyone. Individuals can either accept or reject the policy. Refer to Fig. 20.11.

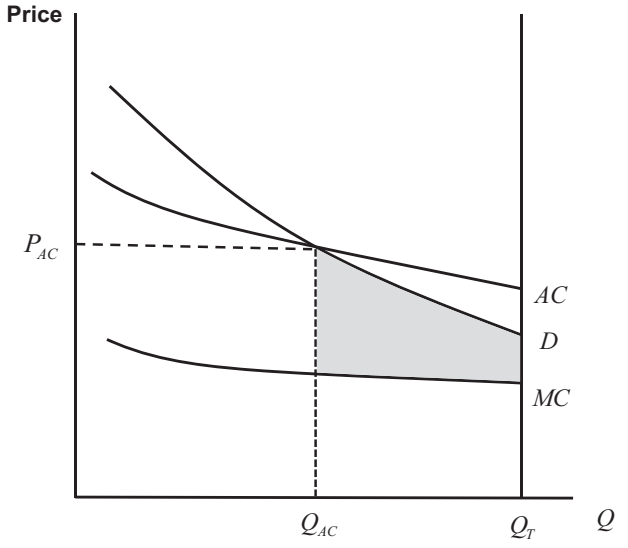

FIGURE 20.11

7. Einav and Finkelstein (2011). The analysis here closely follows their analysis on pp. 1–13.

The price of the policy is on the vertical axis and the number of policies sold is on the horizontal axis. The quantity sold is equal to the number of individuals who accept the insurance, a maximum of Q_T policies. The demand for insurance by each individual depends on two factors, the individual's risk—the probability, π_i, of becoming ill—and his degree of risk aversion. The aversion to risk determines the risk premium he is willing to pay above the actuarially fair premium. The demand curve, D, is downward sloping because the individuals' risks are assumed to decrease from left to right.

The market for insurance has the unusual feature that the demand for insurance is directly linked to the cost of insurance since the π_i represents the marginal cost to the insurer of insuring individual i. The marginal cost curve, MC, in the figure is downward sloping given the ordering of risks. Note that D has to be above MC because D includes the individuals' risk premium as well as their risk. Also, D has a steeper slope than MC because we are also adopting the standard assumption in the literature that risk aversion increases directly with risk. Therefore, the gap between D and MC decreases from left to right.

Since D is everywhere above MC, the efficient outcome is for everyone to be insured. But this does not happen. The insurers' average cost, AC, is above MC because it reflects the average of all the risks insured at any given quantity. Assuming the market is competitive, the break-even price is $P = \text{AC}$, and the number of policies issued is Q_{AC}. The lowest-risk individuals, $(Q_T - Q_{\text{AC}})$, are uninsured even though they want the policy and are willing to pay more than their marginal costs to an insurer to purchase it. The welfare loss is the shaded area between D and MC from Q_{AC} to Q_T (assuming no income effects such that D and D^{comp} are the same). Underinsurance of low-risk types is a common feature of insurance models with adverse selection. Indeed, the term adverse selection derives from the property that the insurers end up with a higher-risk, i.e., more adverse, pool of insured on average than at the efficient outcome because they are forced to charge a single premium equal to average cost.

The standard outcome pictured above does not necessarily occur, however. In Fig. 20.12(a), D is above AC and everyone is insured. This can happen if individuals do not vary much by risk—MC and thus AC are relatively flat, and/or individuals are highly risk averse such that D is well above MC. Unfortunately, the situation pictured in Fig. 20.12(b) is possible as well, in which D is everywhere below AC, perhaps because people are not very risk averse. Suppose insurers, uncertain about an initial price, insure Q_0 individuals at a price P_0. They make a loss and raise the price to AC at Q_0, which lowers the quantity demanded. The average risk of those remaining insured increases, raising AC again and

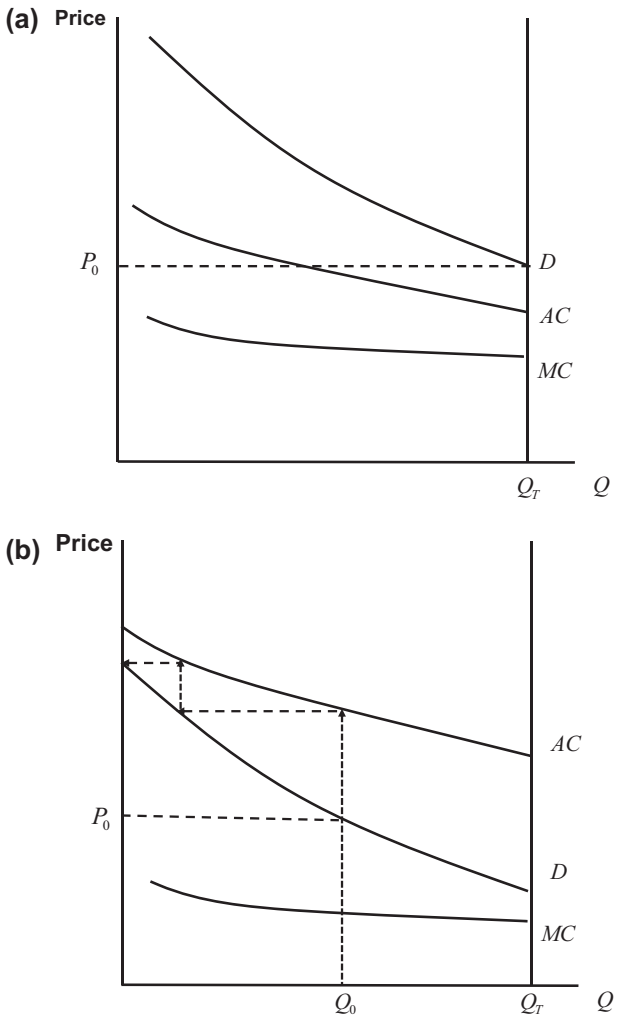

FIGURE 20.12

generating losses, and so forth, following the arrows back. The market completely unravels as the insurance pool becomes ever more adverse, referred to as a death spiral in the insurance literature. Insurance markets can be extremely fragile under adverse selection.

ADVANTAGEOUS SELECTION

Adverse selection and underinsurance are the expected outcomes when insurers cannot differentiate the insured on the basis of risk, but not the only possibility. It is also possible to have advantageous selection and overinsurance. This can happen if risk aversion and the associated risk premiums are inversely related to risk: those with the lowest risks (π_i) have the highest risk premiums. In this case, if the demand curve is drawn as downward sloping, then the marginal cost curve must be upward sloping, since D and MC converge as risks increase and risk premiums decrease. Demand would still be above MC and imply that

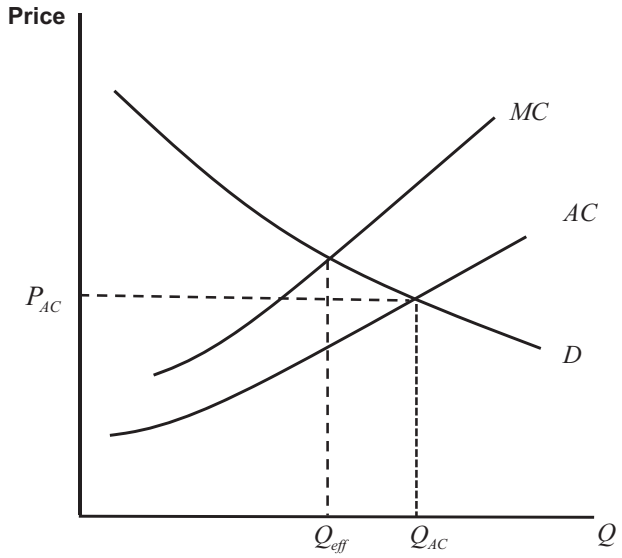

FIGURE 20.13

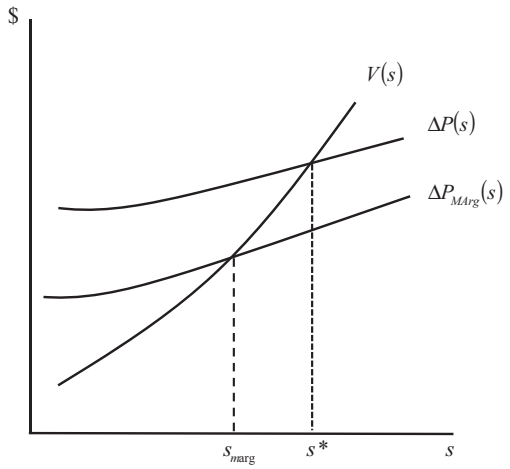

FIGURE 20.14

full insurance is optimal as the model stands so far. This result can be overturned, however, if administrative costs of providing insurance (insurance loads), which we have been ignoring so far, are added to the model. These costs shift up the marginal cost curve and can generate the market situation illustrated in Fig. 20.13.

With MC upward sloping, it lies above AC. The efficient output is Q_{eff}, the intersection of D and MC, but with competition insurers set P = AC, and Q_{AC} individuals purchase the insurance. There are too many people insured. This case is termed *advantageous selection* because the pool of the insured is less risky on average than it would be if the market were efficient. There is some evidence of the existence of advantageous selection in certain markets, but the insurance literature suggests that adverse selection is the usual case.[8] The remainder of the chapter will assume that adverse selection is the problem.

A TWO-POLICY MODEL

Further insights into the issues that arise with adverse selection can be obtained with another simple model by David Cutler, in which individuals can choose between two policies offered by all insurers, a moderate policy and a more generous policy.[9] As above, there is a continuum of individuals differentiated by risk, and the insurers cannot distinguish people by risk. Let s_i be the expected expenditures of individual i under the more generous policy. In Fig. 20.14, the individuals are ordered along the horizontal axis from lowest risk to highest risk, and the line $V(s)$ represents the additional value of the generous plan relative to the moderate plan to an individual with risk s. $V(s)$ is assumed to be upward sloping, because medical expenditures tend to be highly skewed toward the most risky individuals.

Everyone would prefer the more generous plan except that it is more expensive than the moderate plan. Individuals decide which plan to select based on a comparison between the extra value and the extra cost of the more generous plan. Assume that s_G is the price of the more generous plan, equal to the average or expected expenditures of the individuals in the plan. Assume further that expected expenditures in the moderate plan for each individual are just a proportion α of the expected expenditures under the generous plan, $\alpha < 1$. Therefore, the price of the moderate plan is $\alpha\, s_M$, equal to the average of the expected expenditures of the individuals in the moderate plan. The difference in price, ΔP, between the two plans is $s_G - \alpha\, s_M$.

To see the effect of adverse selection on the difference in the prices, add and subtract s_M: $\Delta P = (1 - \alpha)s_M + (s_G - s_M)$. Consider a movement to the right in the figure, which occurs as more people choose the moderate plan. The first term is the cost saving experienced by the average enrollee in the moderate plan. The second term is the effect of adverse selection. As people go to the more moderate plan, the pool of the insured in both plans becomes more risky on average since the least risky individuals in the generous plan are moving to the moderate plan and simultaneously become the most risky individuals in the moderate plan. Because medical expenditures are highly skewed toward the most risky individuals, the average expenditure in the generous plan, s_G, is likely to increase by more than the average

8. E/F note that Finkelstein and McGarry found evidence of advantageous selection in the market for long-term care insurance: people who are less likely to need long-term care are more likely to purchase the insurance (Finkelstein and McGarry, 2006).

9. D. Cutler's chapter is an excellent starting point for a wide range of economic issues in the market for health care, including moral hazard and adverse selection.

expenditure in the moderate plan, s_M. Therefore, $s_G - s_M$ should increase as the number of people in the moderate plan increases and also dominate the first term. Therefore ΔP should rise as more people join the moderate plan, as drawn in the figure.

The individual with risk s^*, at the intersection of ΔP and $V(s)$, is indifferent between the two plans. Everyone with $s > s^*$ buys the generous plan and everyone with $s < s^*$ buys the moderate plan. This is not the efficient outcome, however. Consider the extra cost imposed on the insurer if the marginal person in the moderate plan were to buy the generous plan. The extra cost is $(1 - \alpha)$ $s_{\text{marg}} = \Delta P_{\text{marg}}$, where s_{marg} is the expected expenditures under the generous plan of the marginal person in the moderate plan. ΔP_{marg} is assumed to be less than the extra cost of joining the generous plan, $\Delta P = s_G - \alpha s_M$, as shown in the figure, because medical expenditures are skewed toward the most risky individuals. That is, s_G exceeds s_{marg} by more than αs_{marg} exceeds αs_M. The optimal point of indifference between the two plans is s_{marg} in the figure, at the intersection of $V(s)$ and ΔP_{marg}. Because the price differences in the two plans are based on average risks rather than marginal risks, there are too many individuals with moderate plans and too few with the generous plans relative to the optimum: $s^* > s_{\text{marg}}$. The outcome under adverse selection in this model is more likely than the outcome in the E/F model with only a single policy. Lower-risk individuals are not shut out of the medical insurance market, but they are forced to accept less-comprehensive policies than they would like even though they are willing to pay insurers for their costs of joining more comprehensive plans. The less-risky people are underinsured in this sense.[10]

IS AN EQUILIBRIUM POSSIBLE?

The E/F and Cutler models actually understate the difficulties that arise under adverse selection because of their assumption that the only form of competition among insurers is price competition. The policies offered are assumed to be a given, and competition among the insurers drives prices to the break-even point, to average costs. In fact, insurers tend to compete along both price and coverage dimensions, offering individuals a selection of different policies at different prices. Michael Rothschild and Joseph Stiglitz (R/S) showed that under broader price and coverage competition, insurance markets may not be able to reach an equilibrium (Rothschild and Stiglitz, 1976).

To see the possible outcomes under price and coverage competition, return to the two-person, high-risk, low-risk model that began this section.

10. The generous plan could disappear entirely in this model if the cost effect of adverse selection is so strong that ΔP lies everywhere above $V(s)$.

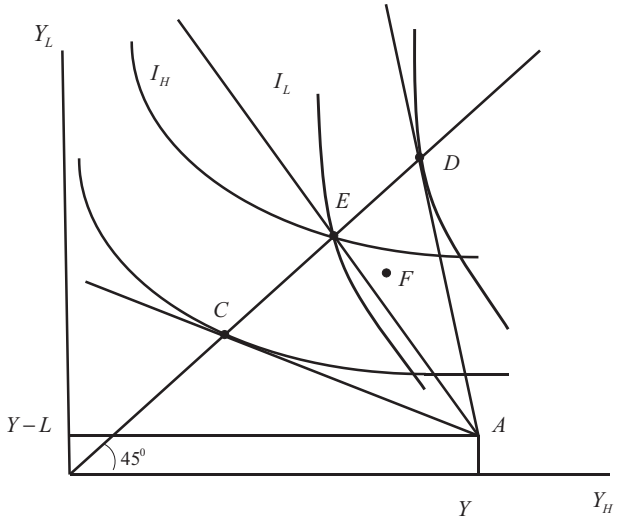

FIGURE 20.15

Figure 20.15 reproduces the equilibria C and D from Fig. 20.10 that would arise if insurers knew who is high- and who is low-risk individual. The high-risk individuals receive actuarially fair, full insurance and move from A to C; the low-risk individuals receive actuarially fair, full insurance and move from A to D. Since the insurers cannot differentiate individuals by risk, they initially offer full insurance to everyone, with both risk groups pooled together. If the proportion of high-risk individuals in the group is λ, then the break-even premium is $P_{\text{avg}} = (1 - \lambda)\pi_L + \lambda\pi_H$, and the fair-odds line has slope $\frac{(1-P_{\text{avg}})}{P_{\text{avg}}}$, line AE in the figure. Both sets of individuals move from A to E, with full insurance at the average premium. E is called a pooling equilibrium. The high-risk individuals are better off than they would be if risks were known to the insurers. The low-risk individuals are worse off. They do not want to pool with the high-risk individuals, as the previous models have suggested, and, with insurers competing in price and coverage, they do not have to. To see why not, notice that at the pooled equilibrium E, the slope of high-risk indifference curve, $\frac{(1-\pi_H)}{\pi_H}$, is flatter than the slope of the low-risk indifference curve, $\frac{(1-\pi_L)}{\pi_L}$. Therefore, insurers can offer partial insurance at a point such as F near E in between the indifference curves, such that the low-risk individuals prefer F to E, whereas the high-risk individuals prefer to remain at E. Moreover, the policy at F is profitable if only the low-risk individuals buy it. It has both less coverage and a lower risk profile than the pooled policy. Since a profitable policy such as F is always possible from a pooled equilibrium, there can never be a stable pooling equilibrium. The only possible stable equilibrium under price and quantity competition is a separating equilibrium, in which each risk class is offered a separate policy.

The problem remains as to the reaction of the high-risk individuals to the introduction of the policy at F. As the low-risk individuals leave the pooled policy at E, the risk pool becomes more and more adverse and the premiums on the policy at E rise. Once they rise, the high-risk individuals move to lower indifference curves and will eventually prefer policy F to the original policy. But once they join F, that policy is no longer profitable. Therefore, a stable separating equilibrium must have the property that the high-risk individuals prefer the policy offered to them to a policy offered to the low-risk individuals. Figure 20.16 illustrates.

The high-risk individuals mostly prefer actuarially fair, full insurance that brings them to point C among all the actuarially fair policies offered only to them. Therefore, having established that the only possible equilibrium is a separating equilibrium, the separating policy tailored to the low-risk individuals must lie on or to the southeast of indifference curve I_0^H on which C lies. Of all possible policies offered to the low-risk individuals, they mostly prefer to be at point G on their fair-odds line AD, purchasing actuarially fair, partial insurance and achieving utility given by indifference curve I_0^L. This is the only possible equilibrium when risks are unknown. The high-risk individuals receive the same—optimal—amount of insurance that they would if their risks were known, whereas the low-risk individuals are forced to accept actuarially fair, partial insurance even though they would prefer full insurance. Ironically, if the high-risk individuals would be willing to reveal their type, then the low-risk individuals would be able to receive actuarially fair, full insurance without any loss to the high-risk individuals.

Unfortunately, even this separating equilibrium may not be stable. Suppose the fair-odds line for a pooled policy is AH, lying above I_0^L. Then a policy offering the combination of premium and coverage at J would be profitable and would be preferred by both high- and low-risk individuals. Therefore, the only possible separating equilibrium would not be stable. Since no pooling equilibrium can be stable either, there is no equilibrium at all in the market. AH can be above I_0^L if either the costs to low-risk individuals of pooling are low or the costs to them of separating are high. The costs of pooling will be low if there are not many high-risk individuals. The costs of separating will be high if they are highly risk averse such that partial insurance leads to large utility losses. Ironically, in the E/F model with a continuum of risks, two conditions under which the efficient outcome of everyone being insured is likely to occur are that the risks are reasonably similar (relatively flat marginal cost curve) and/or the individuals are highly risk averse (high demand curve). Yet these are the conditions for which the no-equilibrium outcome is most likely under price and coverage competition. Indeed, R/S point out that an equilibrium can never exist if there is a continuum of risks under price and coverage competition. The intuition is that there are always individuals close together along the continuum for whom it pays to pool their risks, but, as in the two-person model, a pooling equilibrium is never stable.

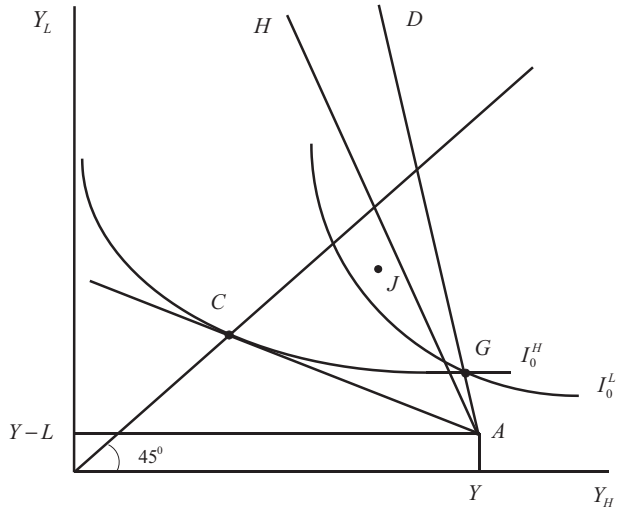

FIGURE 20.16

Conclusions

The models considered in this section yield four conclusions regarding the costs of adverse selection:

a. The insurance market may not have an equilibrium.
b. If an equilibrium does exist, low-risk individuals are likely to be underinsured relative to the coverage they would like to have, and relative to what they would have if insurers had full information about risks, Conversely, high-risk individuals receive actuarially fair, full insurance whether or not their risks are known to the insurers.
c. Insurers have an incentive to offer partial coverage to attract low-risk individuals.
d. Under a separating equilibrium, high-risk individuals pay more for insurance than low-risk individuals; the efficient solution ex ante before risks are known would be a pooling equilibrium in which everyone purchased premium insurance against their eventual risk type and therefore paid the same price for health insurance.

POLICY RESPONSE TO ADVERSE SELECTION

The government's policy response to adverse selection is straightforward. The government should mandate that everyone receive health insurance, either privately or

publicly provided. The mandate is necessary to ensure that low-risk individuals participate in the insurance.

The policy is not without its problems, however. One is that the low-risk individuals will object, since they know they are subsidizing the high-risk individuals. Another is that the poor will require large subsidies if the government forces everyone to pay a premium for the policy.

The government could provide the insurance and raise taxes to cover the costs. Under the U.S. federal personal income tax, most poor have no tax liability. Therefore, tax-financed medical insurance does subsidize the poor. Still, the taxes required to pay for the insurance are themselves distorting and add efficiency losses to the cost of the policy. Finally, a tax-financed government policy could not ignore the potential for ex post moral hazard, as we have in this section. The optimal government policy would almost certainly not be full insurance.

U.S. POLICIES

The three main U.S. public health care initiatives are Medicare, Medicaid and its companion Children's Hospital Insurance Program (CHIP), and the Patient Protection and Affordable Care Act. Medicare and Medicaid were instituted in 1965 as amendments to the Social Security Act, and the Patient Protection and Affordable Care Act was passed in 2010.

MEDICARE AND MEDICAID

Medicare originally provided coverage for hospitalization and physician services for all people aged 65 years and above. It was paid for by placing a part of the payroll tax that finances the Social Security System into a separate Medicare Trust Fund. Hospitalization was offered without premiums to the elderly who were part of the Social Security System and physician services required a small annual premium. The elderly who are not part of the Social Security System can receive Medicare coverage by paying premiums for hospitalization as well as physician services. An option to receive all coverage from private insurers, called Medicare Advantage, was instituted in 1997, and prescription drug coverage was added in 2006.

Medicaid was originally intended to be a fairly modest program. The public assistance programs created by the Social Security Act of 1935 offered both monthly cash payments to recipients and payments to vendors who provided recipients with medical services. The intent of Medicaid was to consolidate all the medical vendor payments under one agency, with some expansion of the medical services offered to the poor. It was subsequently expanded in a series of steps to include nonpoor medically needy families whose incomes were up to three times the federal poverty line. These expansions, combined with the continuing rapid increase in health care costs, drove up Medicaid expenditures to the point that it became larger than all the other federal public assistance programs for low-income families and individuals combined.

Medicaid is administered by the states, which can determine eligibility, coverage, and the income limits below which families are covered, all subject to broad federal guidelines. Families with dependent children and all pregnant women must be covered under Medicaid if their incomes are at or below the federal poverty line.[11] States have the option of expanding coverage to families above the poverty line. The federal government subsidizes the states for their Medicaid expenditures, from 50% to 83%, with the matching rate inversely related to a state's income relative to average state income.

The CHIP, instituted in 1997, covers hospitalization for children in families whose incomes are above the Medicaid maximums in each state and who would otherwise be uninsured. The federal subsidy to states for CHIP is about 15% points higher than for Medicaid.

The motivation for Medicare and Medicaid (including CHIP) was not a matter of overcoming private information. Rather it was distributional, closest in spirit to Nyman's view of granting access to otherwise unaffordable medical care. Medical care came to be viewed as a merit good, one so essential that no citizen should be denied access to it. Regarding Medicare, insurers know that the elderly impose much higher risks than the nonelderly population; 49% of lifetime medical expenditures, on average, are incurred starting at 65 years of age (Alemayehu and Warner, 2004). There is no informational problem of adverse selection. Private insurers would be willing to insure the elderly, but only at very high premiums that many of the elderly could not afford. Similarly, most poor families cannot afford medical insurance. Thus the United States made a collective decision to provide medical insurance to the elderly and the poor, financed by taxes.

The two programs solve the adverse selection problem by including virtually everyone within a particular demographic and/or income category. The ex post moral hazard of overusing medical care remains under these programs, however, and the government tries to reduce the incentives for overuse by means of deductibles and copayments, especially under Medicare. For example, in 2012 Medicare beneficiaries were responsible for the first

11. The Patient Protection and Health Care Reform Act increased mandatory coverage to all individuals in families with incomes less than or equal to 138% of the federal poverty line starting in 2014, and increased the federal subsidy rate to help states cover the additional expenses this would entail.

$1156 of expenditures for days 1 through the 60 of hospitalization, and then daily co-payments for stays beyond 60 days. Recipients also pay $144.50 per day for days 21−100 of rehabilitation in a skilled nursing home. Physician services come with an annual $140 deductible and a co-payment of 20%.[12] Beneficiaries can purchase the so-called Medigap policies from private insurers that cover the Medicare deductibles and co-payments and most do, such that 90% of beneficiaries receive full insurance.[13] Purchasers of Medigap should pay increased premiums to Medicare since they impose more costs on Medicare because of the incentive to overuse medical care with full insurance, but they are not asked to do so. In effect, therefore, Medicare subsidizes the Medigap policies since Medicare pays the majority of the medical costs.

Medicaid is harder to characterize because premiums, deductibles and co-payments, and the income levels below which coverage is offered vary considerably across states. The federal government sets maximum allowable co-payments that are nominal for physicians and drugs, only $3.80 per visit or prescription for those at or below the poverty line. The maximum allowable hospitalization co-payment is not trivial, however: 50% of the costs of the first day of hospitalization for individuals in families at or below the poverty line. This is 50% of the Medicaid payment to the hospital, which is much less than the usual hospital charge to the uninsured. Still, it is a hefty co-payment for the poor. (The maximum allowable co-payments for all services rise with income levels.) No co-payment is required for any medical service incurred by children and pregnant women at or below the federal poverty line, patients who are terminally ill, and patients institutionalized in nursing homes. Therefore, coverage under Medicaid is close to full insurance for many recipients, thereby maximizing the incentives for ex post moral hazard.

An additional problem with Medicaid was discovered when coverage was expanded to low-income families above the poverty line: Medicaid crowded out quite a bit of private insurance. Approximately half of these families canceled medical insurance they had purchased previously and joined Medicaid to save costs.[14] There is no net benefit from consumption smoothing under Medicaid when it substitutes for equivalent private insurance, which substantially lowers the overall benefit of Medicaid to low-income families. Also, the taxes used to finance Medicaid add their own inefficiencies to the costs of Medicaid. Therefore, Medicaid may not pass a cost−benefit test, especially if there is a considerable amount of overuse of medical care given the extent of the coverage. If it does not, then the justification for the program is entirely distributional.

PATIENT PROTECTION AND AFFORDABLE CARE ACT

The Patient Protection and Affordable Care Act was targeted to the remaining nonpoor, non-elderly adult population and their families. The United States is unique among the industrialized market economies in that the medical insurance coverage received by working adults depends on where they work. Employers contract with private insurers to provide group coverage of their employees. Most of the industrialized countries provide some kind of universal coverage for everyone through their governments. One implication of the U.S. system is that people who work for larger companies tend to get more comprehensive and/or cheaper coverage than those who work for smaller companies, which may not provide any coverage for their employees. Another implication is that people lose their insurance coverage if they are laid off or quit their jobs. Non-elderly adults who are not covered by Medicaid and who are either not working or are working in small firms that do not provide medical insurance have to purchase insurance on their own from private insurers. But many do not do so, either because they do not want medical insurance or cannot afford it. Over 50 million people in the United States had no medical insurance in 2010. The main thrust of the Act was to provide coverage for the uninsured.

In the debate leading up to the Act, some people within the Obama administration and Congress wanted to scrap the employer-based system and simply expand Medicare to include all nonpoor adults and their families. This was rejected in favor of retaining the current system, but requiring all noninsured adults to purchase coverage, subject to stiff fines if they refused. Low-income adults are subsidized, so that payments for health insurance are no more than 10% of their gross incomes. The federal government subsidized states to establish health insurance exchanges by 2014, in which the noninsured would buy insurance from private insurers. The idea behind the state-run exchanges was to foster competition among insurers such that the noninsured would have a number of price and coverage options to choose from. Also, employers are subject to very high penalties if they drop the coverage of their employees and force them into the insurance exchanges. If states refused to operate the insurance exchanges, and many did, individuals could purchase the required insurance on an alternative federal insurance exchange.

The Act also prohibits insurance companies from denying individuals coverage on the basis of preexisting medical conditions and from setting lifetime expenditure

12. Klees et al. The document provides an excellent overview of the main provisions of Medicare and Medicare, along with a history of the two programs.
13. Cutler, *op. cit.*, p. 2216.
14. *Ibid.*, p. 2150.

caps on coverage, both common practices among insurers. These provisions took effect in 2014 (although children with preexisting conditions could not be denied coverage by the end of 2010). They were delayed so that the increased cost burden they placed on the insurers would be counterbalanced by the entry into the insurance pools of the previously uninsured, who tend to be younger and healthier than the average adult. The other two major components of the Act are the financing of a variety of pilot programs designed to reduce medical costs, and a number of tax and fee increases to make the Act revenue neutral.

A lawsuit was brought against the constitutionality of the Act, which the U.S. Supreme Court decided to hear. On June 28, 2012, the Court ruled that the mandate and the fee to be paid if a person did not have insurance were both constitutional, because the fee was essentially a tax. This left intact the insurance exchanges and the minimum coverage provisions of all health insurance policies whether offered on the exchanges or not. The decision was not a total victory for the Obama administration, however, since the Court also ruled that states could not be forced to extend their Medicaid coverage to 138% of the poverty line. Many states decided not to extend the Medicaid coverage. This left people between 100% and 138% of the poverty line in limbo, because the subsidies for purchasing health care on the state exchanges began for those families at 138% of the poverty line. Given the political divisions between the Democrats and Republicans at the time, there was no way to amend the law to include health care coverage for those between 100% and 138% of the poverty line.

A question arises about the state-run insurance exchanges based on our analysis of adverse selection: How will the insurance companies respond to the inability to deny coverage on the basis of preexisting conditions or to impose lifetime expenditure caps? Two of the major devices they have been using to limit the riskiness of their insurance pools have been taken away from them. Our analysis suggests that the private insurers will attempt to achieve a separating equilibrium by offering, roughly speaking, two kinds of policies. One is a comprehensive policy with very high premiums targeted to the high-risk individuals among the uninsured. The other is a policy having much more limited coverage and a low premium, with the coverage too limited to attract the high-risk individuals. Presumably some high-risk, low-income individuals will buy the comprehensive policy because their premiums are heavily subsidized by the federal government. But other high-risk individuals may find the premiums too expensive, choosing instead to remain uninsured and pay the fine. The low-risk individuals may purchase the limited policies, but then they are likely to be in the position of receiving much less coverage than they would like and would be willing to pay for. If this kind of separating equilibrium arises, few people are likely to be very happy with the state-run insurance exchanges. One also wonders whether the private insurers can issue profitable policies in the insurance exchanges given their inability to deny coverage based on preexisting medical conditions or to impose lifetime expenditure caps.[15] This will depend in part on the premiums required versus the fines imposed for refusing coverage.

In short, the equilibrium outcome in the state-run insurance exchanges is difficult to predict if, indeed, one exists at all. If many people do turn out to be highly dissatisfied with the exchanges, the proposal for universal coverage provided by the government that was rejected leading up to the Act may become increasingly attractive to the American public. And U.S. firms (other than insurance companies) are unlikely to object to universal government coverage, since they have long complained that the need to offer medical insurance to their employees puts them at a competitive disadvantage in an increasingly international marketplace.

REFERENCES

Alemayehu, B., Warner, K., June 2004. The lifetime distribution of health care costs. Health Services Research 39 (3). Table 4, ncbi.nim.nih.gov/pmc/articles/PMC1361028.

Cutler, D. Health care and the public sector, (Chapter 31). In: Auerbach, A., Feldstein M. (Ed.), Handbook of Public Economics, vol. 4. Elsevier Science B. V., pp. 2199–2201.

Einav, L., Finkelstein, A., January 2011. Selection in insurance markets: theory and empirics in pictures. NBER Working Paper 16723.

Finkelstein, A., McGarry, K., September 2006. Multiple dimensions of private information: evidence from the long-term care insurance market. American Economic Review 96 (5), 938–958.

Klees, B., Wolfe, C., Curtis, C. Brief Summaries of Medicare and Medicaid, Title XVIII and Title XIX of the Social Security Act as of November 1, 2011. Office of the Actuary, Centers for Medicare & Medicaid Services, Department of Health and Human Services, p. 15. Available at: www.cms.gov.

Nyman, J., December 1999a. The economics of moral hazard revisited. Journal of Health Economics 18, 811–824.

Nyman, J., December 1999b. The value of health insurance: the access motive. Journal of Health Economics 18, 141–152.

Pauly, M., June 1968. The economics of moral hazard: comment. American Economic Review 58 (3, Part 1), 531–537.

Pauly, M., February 1974. Overinsurance and public provision of insurance: the roles of moral hazard and adverse selection. Quarterly Journal of Economics 88 (1), 44–62.

Rothschild, M., Stiglitz, J., November 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. Quarterly Journal of Economics 90 (4), 629–649.

15. Adverse selection is not a problem for the employer-based plans because they are enforced pooling equilibria across all the employees within a firm. But if insurance companies have difficulty issuing profitable policies in the state-run insurance exchanges, then they may have to raise premiums on the employer group policies. This may cause some employers to drop their coverage, pay the fines, and force their employees into the exchanges.

Chapter 21

Social Insurance: Social Security

Chapter Outline

A common form of social insurance in the industrialized market economies is the provision of public pensions to protect the standard of living of retirees. Just as in private pension plans, people contribute to the program throughout their working years, usually by means of an earmarked tax payment, and then receive a pension, a stream of income payments during their retirement years. Under some programs, people can elect to receive the income payments once they have reached a certain retirement age designated by the program even if they continue to work. But the main purpose of these programs is to protect the elderly during their retirement years, and we will think of them as public pensions throughout this chapter. Public pensions are commonly called social security.

Social security programs vary along a number of dimensions. To begin with, they can be either defined benefit or defined contribution plans. Under a defined benefit plan, the government provides retirees with income during their retirement years, almost always in the form of an annuity. An annuity is a fixed annual payment that continues until the person dies (married couples can usually choose an option to have the annuity payments continue until the surviving spouse dies as well, in which case the annuity payment each year is reduced). All payments cease at death—there is nothing left to bequeath to heirs. The government determines the annuity payments in defined benefit plans, and they are usually based on the contributions made during the working years. Since employees with higher incomes generally make larger contributions to social security during the working years, the annuity payments are positively related to a person's earnings.

Under a defined contribution plan, an employee's annual contributions are credited to an investment account owned by the employee, called a personal account. The funds in the personal account are then invested in various market financial securities, either by the employee or by the government on the employee's behalf. When the person retires, the entire accumulated account is then made available to him or her. The government may or may not dictate how the funds are used, such as requiring that some or all of the funds be used to purchase an annuity. The more common option, however, is to allow the retirees to use their funds in their personal accounts as they wish.

A second dimension of public pensions is that they can be fully funded, partially funded, or unfunded. Under a fully funded plan, the contributions by all the people during their working years are placed in a fund and invested. The

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00021-2

accumulated assets and the present value of the expected returns to be earned on the assets are just equal at each point in time to the present value of the expected payouts to current and past contributors during their retirement years. A partially funded plan is one in which the expected payouts exceed the accumulated contributions in present value. An unfunded plan is a pay-as-you-go or paygo plan, in which the government collects tax contributions from the current employees and simply pays them out in the same year to the current retirees. There is no accumulation of funds to cover future pension obligations.

A defined contribution plan is necessarily fully funded, since the payouts from the personal accounts can be only what the accumulated funds in each account permit.[1] Defined benefit plans, in contrast, may be fully funded, partially funded, or unfunded, depending on what the government chooses to do with the contributions. We will consider only fully funded or unfunded defined benefit pensions in this chapter. Partially funded pensions add complications that are unnecessary for understanding the basic economic issues associated with social security.

A final distinction worth noting at the outset relates to the contributions to pension plans. The government may mandate given contributions every year, such as by means of an earmarked tax payment. Alternatively, it may offer a subsidy to encourage contributions to a private or public pension plan, without mandating any given amount of contribution. A common choice in the industrialized market economies is a mixture of subsidies to private plans and a mandatory contribution to a public plan.

THE U.S. SOCIAL SECURITY SYSTEM

The U.S. Social Security System was established under the Social Security Act of 1935. President Roosevelt conceived of the Social Security pensions as a defined contribution plan. A new payroll tax of 2%, half paid by employees and half by their employers, provided the contributions to the plan. The contributions would be placed in the Social Security Trust Fund and invested by the government, and the employees would receive an annuity upon retirement based on their accumulated contributions to the Trust Fund. The need to provide immediate income to retirees during the Great Depression undermined Roosevelt's intentions, however, and Social Security quickly turned into an unfunded paygo system. The payroll tax revenues were immediately paid out to covered employees as they retired.

Social Security has essentially remained on a paygo basis, with one notable exception. In 1983, President Reagan and Congress agreed to a set of reforms that were motivated by the huge baby-boom generation born between 1946 and 1964. There were two main reforms. First, Congress had passed legislation in a series of steps to increase the payroll tax. From 2000 on it has been 10.6%. In addition, the retirement age at which people could receive full benefits was increased from 65 to 67 years in increments over time starting with the cohort born in 1938 and ending with the cohort born in 1960. As a result of these reforms, the payroll tax contributions began to substantially exceed the payments to retirees each year and the Trust Fund accumulated a surplus. The surplus was invested in U.S. Treasury securities. The idea was that the accumulated surplus would be sufficient to cover the total expected payments to the baby-boom generation, with the assets in the Trust Fund becoming exhausted once the last baby boomers die. The System would then revert to a paygo system with no further reforms.

The assumptions made at the time turned out to be overly optimistic. Current estimates are that the Trust Fund will be exhausted in 2033, well before the last baby boomers die. Moreover the revenues from the payroll tax every year after 2033 will be insufficient to finance even the current retirees under the existing law.

The post-1983 Trust Fund surplus may have been a bit of a fiction since, by purchasing Treasury securities, it was essentially financing deficits in other parts of the government budget. The important question of how much the 1983 reforms increased saving by the federal government depends on whether the deficits in the other part of the federal budget would have arisen anyway, or were larger than they would have been without the 1983 reforms because of the convenience of the government being able to borrow from the Trust Fund. That is a difficult question to answer.

The Social Security pension system is highly complex. For our purposes, it is sufficient to keep in mind the following four features, in addition to its paygo structure:

Benefits

The benefits that retirees receive are in the form of an annuity. The annual benefit formula is based on the highest-earning years of the retirees, to a maximum of the highest 35 years. Therefore benefits rise with earned income. Nonetheless, the benefit formula is progressive—it gives proportionately more benefit to the lower income earners.

1. The only exception to this are the so-called notional defined contribution plans used by Sweden and Italy, in which the contributions to each individual's personal account are credited with a designated or notional rate of return each year, regardless of what the funds in the account actually earn. At retirement, each person receives the funds accumulated at the notional interest rate. Such accounts are therefore not fully funded in the standard actuarial sense, which assumes that the funds are invested at actual market rates. Indeed, the government may choose to do whatever it wishes with the contributions. It may not accumulate any fund at all.

That is, Social Security is not a straight insurance plan; it intentionally redistributes within each generation. Finally, the annual benefits are indexed to inflation (the Consumer Price Index).

Spousal Benefits

In two-earner families, the spouse with the lower income receives a maximum benefit equal to one-half of the benefit of the higher earner. The lower earner was typically the wife throughout the twentieth century.

Payroll Tax Contributions

The payroll tax to finance the benefits is levied at a single rate on wage income only, and then only to a maximum amount of income—$106,000 in 2012. The income maximum is also indexed to inflation. Employees and their employers split the tax rate, each contributing half of the revenues for each employee. The combined tax rate earmarked for the pensions was 10.6% in 2012.

Coverage

Coverage expanded enormously throughout the twentieth century, from 35% of all workers in 1935 to 96% of all workers in 2000 (Feldstein and Liebman, 2002). Under the paygo feature prior to 1983, current retirees received much larger payments every year, commensurate with the increasing payroll tax collections, such that the pre-1983 generations of retirees received huge returns on their payroll tax contributions made during their working years. Congress was willing to increase retirement benefits to provide a more reasonable standard of living for retirees, who at that time had higher-than-average poverty rates.[2]

SOCIAL SECURITY AS SOCIAL INSURANCE

Two natural questions arise regarding social security: Why should governments require people to save for their retirement? and Why force retirees to accept an annuity in return for the required saving, as the U.S. Social Security System does?

A decision by society to force people to save for their retirement is typically justified on paternalistic and distributional grounds. Both were clearly motivating factors behind the Social Security Act of 1935. President Roosevelt wanted to ensure that people would save adequately for their retirement and not become wards of the state. Many retirees were in desperate straights at the time. It does not require a Great Depression, however, to appreciate that many people will not, and perhaps cannot, save adequately for their retirement. Planning optimally, or even reasonably well, for retirement is a decision process with a very long time horizon that requires frequent reoptimizing along the way as economic circumstances change. This is likely to be extremely difficult for all but the more financially astute individuals. Small wonder that Alicia Munnell found in 2004 was that the median financial wealth of heads of U.S. households near retirement (ages 55–64 years) was only $30,000 (Munnell and Sunden, 2006). The argument for social security on paternalistic grounds is hardly unreasonable in the United States.

The second question of why social security benefits should be in the form of an annuity is based on standard social insurance efficiency arguments. There is a long-established market for private annuities in the United States (and elsewhere). This is hardly surprising, since the purchase of an annuity is the cheapest way for retirees to provide, ex ante, a given stream of income for themselves for the rest of their lives. Compare, for example, the choice of purchasing an annuity versus living off the principal and interest of bonds to provide the same annual income stream for retirement. Insurance companies can offer annuities more cheaply than the cost of the bonds precisely because the annuity payments cease at death, whereas the bonds pay interest until they mature regardless of whether the bondholder lives or dies until the maturity date. Even if people want to leave a bequest, they should still purchase an annuity to provide for their own consumption during retirement. This will permit a larger expected bequest from any funds remaining, with the further advantage that the expected value of the bequest is independent of how long one lives. These advantages of private annuities as a retirement asset notwithstanding, a standard social insurance argument can be made to justify the provision of public annuities.

There are two uncertainties associated with retirement: the date of retirement and the date of death. Moreover they both exhibit private or asymmetric information, since individuals are likely to have a better sense of each of them than any insurance company would have. As such, they can give rise to the twin problems of adverse selection and moral hazard.

The problem of adverse selection for private annuities is the opposite of that for health insurance: Insurers would like to sell annuities to people who are in poor health, with short life expectancies, but annuities are more attractive to relatively healthy people who are expected to live for a long time. The result is that too many people with long life expectancies buy private annuities, which drives up their

2. For details on the Social Security System, consult the Annual Reports of the Trustees of the Federal Old Age and Survivors Insurance and Federal Disability Insurance Trust Funds. The Annual Reports include 75-year projections of Social Security expenditures and revenues. The 2012 Annual Report is available on the Social Security Web site at www.ssa.gov/oact/tr/2012/index.html.

price and makes them even less attractive to people with short life expectancies. The market is inefficient, as described in the previous chapter, and can even unravel entirely if the pool of insured becomes evermore adverse.

The problem of moral hazard arises if the annuities are truly retirement annuities with payments beginning at the date of retirement, as are private pension plans. People can choose to retire early to trigger the payments, increasing the cost of the annuities. In the U.S. Social Security System, the offer to let people retire early beginning at 62 years of age with reduced benefits is a potential source of moral hazard.

Another problem with private annuities is that insurers cannot write policies that adequately protect retirees for inflation, since inflation is a systemic rather than an individual risk and therefore uninsurable. Some company pension plans do provide inflation protection, but it is almost always limited, with a fairly low ceiling of 2–3% per year.

The market for private annuities is extremely thin in the United States and has very high administrative costs, 10% versus 6% for private pension plans and 0.6% for Social Security. The market is so thin that Peter Diamond believes people must not understand the advantages that annuities have over other forms of retirement investments. This leads him to support Social Security's defined benefit annuities for paternalistic as well as the standard efficiency reasons, in addition to supporting social security programs generally on paternalistic grounds because of the difficulties of saving adequately for retirement (Diamond, 2004).

THE MACROECONOMIC EFFECTS OF SOCIAL SECURITY

Social security systems have a number of macroeconomic effects that are absent in other forms of social insurance programs such as health insurance. Of particular importance is the effect of social security on the rate of saving in the economy, and thus on the rate of investment and long-run economic growth. Another effect of note is the possibility that social security can cause a substantial amount of redistribution across generations. Both effects depend on the structure of the social security system, whether it is a defined benefit or defined contribution program and, if the former, whether it is fully funded, partially funded, or unfunded. We begin with the effect of social security on the rate of saving.

SOCIAL SECURITY AND SAVING

The Samuelson Consumption-Loan Model

In 1958, Paul Samuelson provided the first formal economic justification for an unfunded paygo social security system (Samuelson, 1958). His model was the by-now standard, two-period overlapping generations (OLG) model consisting of a young generation of workers and an old generation of retirees. Individuals are identical and each wants to maximize a two-period additively separable utility function whose arguments are consumption each period, c_{1t} and c_{2t+1}.

$$U = U(c_{lt}) + (1+\partial)^{-1}U(c_{2t+1}) \qquad (21.1)$$

∂ is the discount rate that the individuals apply to future consumption, their rate of time preference.

Each person supplies a fixed amount of labor, l_t when young and receives the competitive wage w_t. We will normalize $l_t = 1$. The consumption good is produced with the labor supplied by the young each period according to the aggregate constant returns to scale (CRS) production function $C_t = aL_t$, where a = the marginal product of labor $= w_t$, and L_t is the number of young workers.

The problem individuals face is that there is no capital in the economy, which implies that the consumer good is perishable and lasts only one period. Otherwise it would act as a capital good if it were storable. Therefore, the market offers no opportunity for individuals to save for their retirement. They would have to work both periods to survive when old if, indeed, the elderly are still able to work.

Samuelson showed that an unfunded social security program can provide for everyone's retirement years so long as the population is growing over time. In the first time period, t_0, the young workers transfer a portion of their income, d_0, to the elderly, which is equivalent to transferring d_0 of consumption to the elderly. To mimic the paygo U.S. system, think of d_0 as the revenue from a payroll tax levied at a rate of τ on w, with $d_0 = \tau\, w$, and with the revenues transferred immediately to the elderly to finance their consumption in retirement. The total amount transferred is d_0L_0. In the next period t_1, the young workers transfer the same portion of their income, $d_1(=d_0)$, to the elderly, who were the young generation in t_0. If the population is growing each year at rate n, then $L_1 = (1+n)L_0$. Therefore, the total amount transferred to the elderly, d_1L_1, is larger by $(1+n)$ than the amount that the current elderly paid to the previous elderly when they were young: $d_1L_1 = (1+n)d_0L_0$. The current elderly receive a rate of return of n on their contribution in the previous period, which Samuelson referred to as the biological rate of return. The pattern repeats itself each period, with each cohort receiving a rate of return n on their contribution to social security when young to finance their retirement. Every generation gains so long as the population continues to grow and the economy never ends; the unfunded social security system represents a pareto improvement in a capital-less economy.

An important extension of the Samuelson model is to allow for the possibility of productivity growth. Suppose

that labor productivity is growing at a rate g per year, such that the aggregate production function becomes $C_t = a(1 + g)^t L_t$. With the wage equal to labor's marginal product, $w_t = a(1 + g)^t$ and incomes also grow at rate g over time. Let the contribution, d_t, that each generation makes to the unfunded social security system represent a constant proportion of each individual's income over time, as it would be under a single-rate payroll tax, rather than a constant dollar amount. Since income is growing at rate g over time, the contribution that each young person makes also grows at rate g over time. Therefore, the transfers received when elderly are greater than the contributions made when young by the factor $(1 + g)(1 + n)$. The rate of return on the contributions made when young is approximately $g + n$, the rate of productivity growth plus the rate of population growth (ignoring the interaction term gn). Under a single-rate payroll tax, this is the annual rate of growth of total wage income, the tax base. It is also the rate of growth of the economy in this model.

Adding Capital to the OLG Model: The Diamond Model

In 1965, Peter Diamond extended the Samuelson OLG model by including a capital market. Diamond's model quickly became the model of choice for analyzing public sector issues in a dynamic context, such as the burden of the public debt, which was the focus of Diamond's paper, and the consequences of different kinds of social security systems, the focus of this chapter.[3]

Adding a capital market, and therefore the possibility of saving for retirement, has a number of important implications for social security, especially for the unfunded system that Samuelson analyzed. An unfunded social security system is no longer pareto improving, in general, in two respects. It can be expected to lower the rate of saving in the economy, thereby reducing long-run economic growth and consumption per capita. Also, some generations gain and others lose from the annual taxes and transfers. Another important implication is that a defined contribution system of personal accounts becomes possible when people are able to save. As we will see, a defined contribution play dominates a paygo plan in an OLG model. It does not reduce saving and there can be no intergenerational redistributions with contributions to personal accounts.

3. Diamond (1965). Our analysis of the Diamond model and the effects of social security closely follow the presentation in Blanchard and Fisher (1989), Chapter 3, pp. 91–104 and 110–113. The main difference in the Diamond and Blanchard/Fisher analysis is that Blanchard/Fisher assumed additively separable utility over time, whereas Diamond used the more general utility formulation, $U = U(c_{1t}, c_{2t+1})$.

The addition of a capital market to the OLG model is hardly a trivial extension. We need to understand the analytics of such an economy before considering the effects of different kinds of social security systems.

THE STRUCTURE OF THE ECONOMY: CONSUMPTION, PRODUCTION, AND MARKET CLEARANCE

Consumption

We begin with the baseline model in which individuals have to provide entirely for their own retirement through their first-period saving—there is no public pension program. As in the Samuelson model, consumers want to maximize a two-period additively separable utility function $U = U(c_{1t}) + (1 + \partial)^{-1} U(c_{2t+1})$. The maximization is subject to two constraints, one in each period. In the first period, $c_{1t} + s_t = w_t$; consumers decide how much to consume and save out of their fixed wage income. As in the Samuelson model, the supply of labor is fixed and normalized to 1. In the second period, $c_{2t+1} = (1 + r_{t+1})s_t$; second-period consumption equals the savings in the first period plus the return, r_{t+1}, on the saving, assumed to be paid in the second period. As above, the economy is perfectly competitive and the consumers take w_t and r_{t+1} as given. There are no taxes on wages or capital in this baseline model. Therefore, r_{t+1} is the marginal product of capital.

Therefore, each (identical) consumer's problem is to

$$\begin{aligned} &\text{Max } U = U(c_{1t}) + (1 + \partial)^{-1} U(c_{2t+1}) \\ &\{c_{1t},\ c_{2t+1}\} \\ &\text{s.t.} \\ &c_{1t} + s_t = w_t \\ &c_{2t+1} = (1 + r_{t+1})s_t. \end{aligned}$$

Eliminate s_t by substituting its value from the first constraint, $s_t = w_t - c_{1t}$, into the second constraint. There is now a single lifetime budget constraint, $c_{2t+1} = (1 + r_{t+1})(w_t - c_{1t})$.

Forming the Lagrangian equation,

$$\begin{aligned} \text{Max } L = U &= U(c_{1t}) + (1 + \partial)^{-1} U(c_{2t+1}) \\ &\quad + \lambda_1((1 + r_{t+1})(w_t - c_{1t}) - c_{2t+1}) \\ \{c_{1t}, c_{2t+1}\} & \end{aligned}$$

The FOC are

$$c_{1t}: \quad U'_1 - \lambda_1(1 + r_{t+1}) = 0 \tag{21.2}$$

$$c_{2t+1}: \quad (1 + \partial)^{-1} U'_2 - \lambda_1 = 0 \tag{21.3}$$

Therefore,

$$U'_1 = (1 + \partial)^{-1}(1 + r_{t+1}) U'_2 \tag{21.4}$$

Consumers equalize the marginal utility of consumption over the two periods, taking into account their rate of time preference and the market rate of return on their saving.

The FOC, combined with the lifetime budget constraint, can be solved for a saving function whose arguments are w_t and r_{t+1}: $s_t = s_t(w_t, r_{t+1})$. The derivative of s with respect to w, s_w, is assumed to be positive under the assumption that both c_{1t} and c_{2t+1} are normal goods. The sign of the derivative of s with respect to r, s_r, is ambiguous, however. An increase in r_{t+1} lowers the price of future consumption relative to current consumption and therefore generates a substitution effect that favors future consumption. Less current consumption implies more saving; the substitution effect makes s_r positive. But an increase in r_{t+1} also relaxes the lifetime budget constraint, and the increase in purchasing power generates an income effect that favors both current and future consumption. More current consumption implies less saving; the income effect makes s_r negative. Therefore, the sign of s_r depends on the relative strength of the substitution and income effects, an empirical question.

Production

Aggregate output is produced each period by a CRS production function whose arguments are the total amount of capital and labor in the economy, $Q_t = F(K_t \cdot L_t)$. The supply of labor, L_t, is fixed and equal to the number of young workers at time t. The supply of capital, K_t, depends on the total amount of saving by the young in period $t-1$. Q can be used either as a consumption good or a capital good and serves as the numeraire, with a price of one. Because production is CRS, the production function can be expressed in per capita terms: $(Q_t/L_t) = F(K_t/L_t, 1)$ or $q_t = f(k_t)$. $f(k_t)$ is concave, with $f' > 0$, and $f'' < 0$ and satisfies the Inada conditions: $f'(0) \to \infty$, $f'(\infty) \to 0$.

With perfectly competitive markets for labor and capital and CRS production:

$$f'_t = r_{t+1} \tag{21.5}$$

$$w_t = f(k_t) - k_t f'(k_t) \tag{21.6}$$

Market Clearance

The total output available at time t is $F(K_t, L_t)$, the output produced, plus K_t, the capital at time t which was the result of the saving in time $t-1$. The output can be used either as capital to be carried forward to time $t+1$, K_{t+1}, or as consumption by the young and the elderly at time t. Letting L_t equal the number of young workers at time t, total consumption at time t is $C_t = L_t c_{1t} + L_{t-1} c_{2t}$. Therefore, the market clearance equation is

$$K_t + F(K_t, L_t) = K_{t+1} + L_t c_{1t} + L_{t-1} c_{2t} \tag{21.7}$$

Dividing Eqn (21.7) by L_t to express market clearance in per capita terms, and recalling that the population is growing at rate n,[4]

$$k_t + f(k_t) = (1+n)k_{t+1} + c_{1t} + (1+n)^{-1} c_{2t} \tag{21.8}$$

The elderly consume k_t, which was the result of their saving in time $t-1$ when young, plus interest on the capital, equal to $k_t f'(k_t)$. The remaining output on the LHS of Eqn (21.8) is $f(k_t) - k_t f'(k_t)$, the wage income of the young workers. Their wage income minus their consumption, $f(k_t) - k_t f'(k_t) - c_{1t}$ is their saving, s_t, equal to $(1+n)k_{t+1}$, the capital available for the next period. Therefore, Eqn (21.8) implies

$$s_t = (1+n)k_{t+1} \tag{21.9}$$

Equations (21.4), (21.5), (21.6), and (21.8) or (21.9) provide a complete description of the operation of the economy.

Dynamics (a): The Steady State

Equation (21.9) is called the accumulation equation because it describes the evolution of the capital stock over time. Recall that $s_t = s_t(w_t, r_{t+1})$. But $w_t = f(k_t) - k_t f'(k_t)$ and $r_{t+1} = f'(k_{t+1})$ from the assumption of perfectly competitive capital and labor markets. Therefore, $s_t(w_t, r_{t+1}) = s_t(f(k_t) - k_t f'(k_t), f'(k_{t+1}))$, and

$$(1+n)k_{t+1} = s_t(f(k_t) - k_t f'(k_t), f'(k_{t+1})) \tag{21.10}$$

an equation in k_t and k_{t+1}. Totally differentiating Eqn (21.10) with respect to k_{t+1} and k_t describes the dynamics of the capital stock over time:

$$\mathrm{d}k_{t+1}(1 + n - s_r(k_{t+1})f''(k_{t+1})) = \mathrm{d}k_t(s_w(k_t) - k_t f''(k_t)) \tag{21.11}$$

or

$$\frac{\mathrm{d}k_{t+1}}{\mathrm{d}k_t} = \frac{-s_w(k_t)k_t f''(k_t)}{1 + n - s_r(k_{t+1})f''(k_{t+1})} \tag{21.12}$$

Equation (21.12), which describes the evolution of the capital stock over time, is entirely determined by individuals' saving behavior. In standard OLG models, such as this, individuals are assumed to be life-cycle savers.

Notice that the sign of Eqn (21.12) is ambiguous. The numerator is positive, but the sign of the denominator depends on the sign of s_r. It is positive if s_r is positive (the substitution effect of the rate of return on saving dominates the income effect); it can be either positive or negative if s_r is negative (the income effect dominates the substitution effect).

4. $k_{t+1} = (K_{t+1}/L_{t+1})$ and $L_t = (1+n)^{-1} L_{t+1}$. Therefore, $(K_{t+1}/L_t) = (1+n)k_{t+1}$. Similarly, $(L_{t-1}/L_t)c_{2t} = (1+n)^{-1} c_{2t}$.

We can sign s_r by making use of the correspondence principle. For the dynamics of the model to be sensible, dk_{t+1}/dk_t must be greater than zero so that the capital stock is either constantly increasing or decreasing from some initial level. In addition, the system must reach an equilibrium steady state in which k is eventually constant. These conditions require that $0 < \frac{dk_{t+1}}{dk_t} < 1$ or, equivalently $0 < \frac{-s_w(k_t)k_t f''(k_t)}{1+n-s_r(k_{t+1})f''(k_{t+1})} < 1$. This is possible at all values of k only if $s_r > 0$ (the substitution effect dominates), which we will assume throughout the remainder of the chapter.

Figure 21.1 illustrates the dynamic behavior of the capital stock. k_{t+1} is on the vertical axis and k_t is on the horizontal axis. The 45°-line represents the steady state in which $k_{t+1} = k_t$. The concave line is the saving function $s_t(f(k_t) - k_t f'(k_t), f'(k_{t+1}))/(1 + n)$. Starting from k_0, the economy moves over time as indicated by the arrows until the saving function intersects the 45°-line at the steady state capital stock, k^*. There is no guarantee that the saving function will be concave as drawn, however. It could increase and decrease in various regions, leading to multiple possible steady states. Or it could be everywhere below the 45°-line, in which case there is no steady state equilibrium. Nonetheless, we will assume throughout the chapter that the dynamics are as drawn and generate a unique steady state.

Dynamics (b): The Golden Rule and Dynamic Efficiency

The accumulation equation also describes the relationship between consumption and the capital stock in the steady state. Rewrite Eqn (21.8), the accumulation equation in terms of output and consumption:

$$k_t + f(k_t) = (1+n)k_{t+1} + c_{1t} + (1+n)^{-1}c_{2t} \quad (21.13)$$

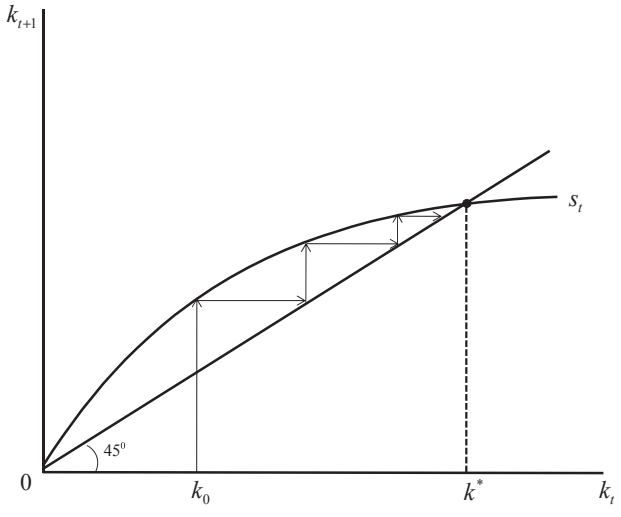

FIGURE 21.1

let $c_t = c_{1t} + (1 + n)^{-1}c_{2t}$, the aggregate consumption per capita at time t. In the steady state, both the capital stock and aggregate consumption are constant: $k_t = k_{t+1} = k^*$ and $c_t = c^*$. Therefore, the accumulation equation in the steady state is

$$k^* + f(k^*) = (1+n)k^* + c^*. \quad (21.14)$$

Rearranging terms:

$$f(k^*) - nk^* = c^* \quad (21.15)$$

Presumably the goal of the economy is to maximize aggregate steady state consumption per capita. This is achieved at the steady state capital stock for which $\frac{dc^*}{dk^*} = 0$. From Eqn (21.15),

$$\frac{dc^*}{dk^*} = f'(k^*) - n = 0 \quad (21.16)$$

Therefore, steady state consumption per capita is maximized when $f'(k^*) = n$, the marginal product of capital equals the rate of growth of the population. Condition (21.16) is referred to as the Golden Rule of Capital Accumulation. If the model were to allow for annual productivity growth at rate g, it turns out that the Golden Rule would be $f'(k^*) = n + g$. In general, steady state consumption per capita is maximized when the marginal product of capital equals the long-run rate of growth of the economy.

The concept of the dynamic efficiency/inefficiency of the economy refers to the inability/ability to increase consumption per capita for all generations, in line with the static concept of pareto optimality. The most common variation of dynamic efficiency refers to the trade-off between current and steady state consumption. If it is possible to increase consumption per capita in the current time period only by decreasing consumption in the steady state, then the economy is said to be dynamically efficient. If it is possible to increase the consumption in both the current time period and the steady state, then the economy is said to be dynamically inefficient. Whether the economy is dynamically efficient or inefficient depends on the steady state accumulation equation.

Suppose the economy has achieved its steady state and there is an increase in consumption. Since output can only be used for consumption or capital, there is an immediate decrease in the capital stock. Suppose the decrease in capital is maintained over time. From Eqn (21.16), $\frac{dc^*}{dk^*} >,< 0$ depending on whether $f'(k^*) >,< n$. If $f'(k^*) > n$ (or, more generally, the rate of growth of the economy), then $\frac{dc^*}{dk^*} > 0$. The decrease in the capital stock reduces steady state consumption. The young and old in the current time period gain at the expense of the future generations; the economy is dynamically efficient. If $f'(k^*) < n$ (or, more generally, the rate of growth of the economy), then $\frac{dc^*}{dk^*} < 0$. The decrease in the capital stock increases steady

state consumption and all generations gain; the economy is dynamically inefficient. In driving the marginal product of capital below the rate of growth of the economy, society is devoting so many of its resources to capital that it reduces consumption possibilities for everyone.

Refer back to Fig. 21.1. Suppose at k^* it happened that $f'(k^*) = n$, the Golden Rule at which c^* is maximized. An increase in saving (decrease in consumption) would shift the saving line up and generate a steady state with a higher capital stock. The economy would be in the dynamically inefficient range because n remains constant but f' would decrease at the higher capital stock. Conversely, reducing saving (increasing consumption) would generate a steady state with a lower capital stock. Steady state consumption would necessarily drop, but since f' would be greater than n, the economy would be in the dynamically efficient range.

We will assume throughout the remainder of the chapter that the economy is dynamically efficient. This is a reasonable assumption for the United States. The marginal product, f', refers to the before-tax rate of return to capital, which is likely to be well above 5% in the United States. Given that annual population growth is only about 1% and productivity growth is on the order of 2−3% per year, the maximum feasible long-run growth of the economy is only 3−4% per year. Therefore, $f' > n$; the capital stock is below the Golden Rule capital stock. Consumption per capita is not maximized but the economy is dynamically efficient.

The Social Welfare Optimum

The final preliminary exercise is to consider the social welfare optimum for the economy. The typical assumption is that the current young and old generations, who determine the social welfare function, care about all future generations, but not as much as they care about themselves. Moreover, they care proportionately less about future generations the farther in the future they are. Therefore, the individualistic intertemporal social welfare function is represented as a discounted sum of the utilities of each generation, with the discount rate ρ representing the social rate of time preference. In terms of our model, if the current generations at time 0 care about themselves and all future generations at time T, the social welfare function is

$$W = (1+\delta)^{-1}U(C_{20}) + \sum_{t=0}^{T}(1+\rho)^{-(t+1)}(U(c_{1t}) + (1+\delta)^{-1}U(c_{2t+1})) \tag{21.17}$$

The first term is the current older generation at time 0. The first term in the summation is the lifetime utility of the current young generation, which in this formulation is discounted by $(1 + \rho)$ because the young generations live for one more year into the future.

A question on which economists have not reached a consensus is what should be the value for the social rate of discount ρ. The most common assumption is that it is a small positive number, such as 1% or 2%. But even a small social rate of discount amounts to essentially ignoring unborn generations who appear far into the future. Another suggestion is to give equal weight to all generations, in which case $\rho = 0$. But this in effect leads to a bias of transferring resources to individuals in future generations because the number of people in each future generation increases at the rate n. Treating all individuals equally through time in the spirit of the Benthamite utilitarian social welfare function would weight each generation by $(1 + n)^{-1}$, such that the social rate of discount would be negative: $(1 + \rho) = (1 + n)^{-1}$. We will follow the usual convention and assume that the social rate of discount is a small positive number, say 1% ($\rho = 0.01$).

The goal is to maximize social welfare with respect to c_{it}, c_{2t}, and k_t, $t = 1, \ldots, T$, subject to the resource constraint, which, as we saw above, is the accumulation Eqn (21.8) in each time period t: $k_t + f(k_t) = (1 + n)k_{t+1} + c_{1t} + (1 + n)^{-1}c_{2t}$. Using the accumulation equation to substitute for c_{1t} in the social welfare function leads to the unconstrained maximization problem[5]

$$\begin{aligned}\underset{(c_{2t},k_t)}{\text{Max}}\ W = {} & U(k_{t-1} + f(k_{t-1}) - (1+n)k_t \\ & - (1+n)^{-1}c_{2t-1}) + (1+\delta)^{-1}U(c_{2t}) \\ & + (1+\rho)^{-1}\left[U(k_t + f(k_t) - (1+n)k_{t+1}\right. \\ & \left. - (1+n)^{-1}c_{2t}) + (1+\delta)^{-1}U(c_{2t+1}) + \ldots\right.\end{aligned}$$

The FOC are

$$c_{2t}: \quad (1+\delta)^{-1}U'(c_{2t}) - (1+\rho)^{-1}(1+n)^{-1}U'(c_{1t}) = 0 \tag{21.18}$$

$$k_t: \quad -(1+n)U'(c_{1t-1}) + (1+\rho)^{-1}(1+f'(k_t))U'(c_{1t}) = 0 \tag{21.19}$$

$$t = 1, \ldots, T$$

The first equation gives the optimal allocation of consumption between the current young and old in each year. The second equation gives the optimal intertemporal allocation for the young across generations. Reducing consumption by the young now generates more capital, which increases the stock of capital and therefore the

5. Two additional constraints that must be specified are the initial and final values of the capital stock, k_0 and k_{T+1}. We assume $k_{T+1} = 0$, since $T + 1$ is beyond the planning horizon.

consumption of the young next period. The second equation equates these marginal losses and gains in consumption, with future consumption discounted by the social rate of discount.

Two more results should be noted. First, combine the two FOCs by solving for $U'(c_{1t})$ in condition (21.18) and substituting the result for $U'(c_{1t})$ in condition (21.19). From condition (21.18):

$$U'(c_{1t}) = (1+\rho)(1+n)(1+\delta)^{-1}U'(c_{2t}). \quad (21.20)$$

Substituting for $U'(c_{1t})$ in condition (21.19) yields

$$-U'(c_{1t-1}) + (1+\delta)^{-1}\left(1+f'\left(k_t\right)\right)U'\left(c_{2t}\right) = 0. \quad (21.21)$$

But $f'(k_t) = r_{t+1}$, the market rate of return. Therefore,

$$-U'\left(c_{1t-1}\right) + (1+\delta)^{-1}(1+r_{t+1})U'(c_{2t}) = 0. \quad (21.22)$$

Equation (21.22) is the allocation rule that each individual uses to divide consumption between the two periods of life, Eqn (21.4) above. It is not surprising that an individualistic social welfare function honors the individuals' intertemporal allocation rule.

Second, the social welfare optimum implies a modified Golden Rule of Capital Accumulation. In the steady state, $c_{1t-1} = c_{1t} = c_1^*$. Therefore, condition (21.19) in the steady state is

$$-(1+n) + (1+\rho)^{-1}(1+f'(k^*)) = 0, \quad (21.23)$$

or

$$(1+f'(k^*) = (1+n)(1+\rho). \quad (21.24)$$

Therefore, $f'(k^*) = n + \rho$ (approximately). The Modified Golden Rule says that the marginal product of capital should equal the rate of growth of the population plus the social rate of time preference (more generally, allowing for productivity increases, the rate of growth in the economy plus the social rate of time preference). It is no longer optimal to maximize consumption per capita in the steady state because the social welfare function applies an ever-increasing discount factor to future generations. There is a trade-off between efficiency and intertemporal equity that requires higher current consumption, a lower stock of capital, and lower consumption per capita in the steady state.

Saving with an Unfunded Social Security System

The pareto improvement that an unfunded social security system brings to Samuelson's model without capital does not apply in a more realistic setting with capital. To the contrary, an unfunded system can generate considerable harm by reducing the rate of saving in the economy.

To see the effect of an unfunded system on saving, return to the consumer's first-order condition (21.4) in our baseline model, $U'_1(c_{1t}) = (1+\partial)^{-1}(1+r_{t+1})U'_2(c_{2t+1})$. Rewrite the condition in terms of saving, with $c_{1t} = w_t - s_t$ and $c_{2t+1} = s_t(1+r_{t+1})$:

$$U'_1(w_t - s_t) = (1+\partial)^{-1}(1+r_{t+1})U'_2(s_t(1+r_{t+1})) \quad (21.25)$$

In an unfunded social security system, each individual pays an amount d_t when young and receives a payment when old of $(1+n)d_{t+1}$, where d_{t+1} is the payment of each young person in period $t+1$. Assume that $d_t = d_{t+1}$, as would be the case with a single-rate payroll tax and no productivity growth, such that $w_t = w_{t+1}$. The consumer's FOC (21.25) becomes

$$U'_1(w_t - s_t - d_t) = (1+\partial)^{-1}(1+r_{t+1})U'_2(s_t(1+r_{t+1}) + (1+n)d_{t+1}) \quad (21.26)$$

To see the effect of the unfunded system on saving, differentiate Eqn (21.26) with respect to s_t and d_t $(=d_{t+1})$, holding constant w and r.

$$\delta s_t(-U''_1 - (1+\delta)^{-1}(1+r_{t+1})(1+r_{t+1})U''_2) + \delta d_t(-U''_1 - (1+\delta)^{-1}(1+r_{t+1})(1+n)U''_2) = 0 \quad (21.27)$$

$$\frac{\partial s_t}{\partial d_t} = -\frac{U''_1 + (1+\delta)^{-1}(1+r_{t+1})(1+n)U''_2}{U''_1 + (1+\delta)^{-1}(1+r_{t+1})(1+r_{t+1})U''_2} \quad (21.28)$$

Both the numerator and denominator are negative with concave utility. Therefore $\frac{\partial s_t}{\partial d_t} < 0$; an unfunded social security system reduces private saving. This in turn implies that national saving (private plus government saving) in the economy decreases since the unfunded system is a pure tax and transfer from young to old that has no effect on government saving. In addition $\left|\frac{\partial s_t}{\partial d_t}\right| <, =, > 1$ as $n <, =, > r$. If the economy is dynamically efficient, $n < r$ and $\left|\frac{\partial s_t}{\partial d_t}\right| < 1$. There is less than complete crowding out of saving by the payment to the social security system.

The effect on the dynamics of the economy is given by the accumulation equation expressed in terms of saving: $(1+n)k_{t+1} = s(w_t(k_t), r_{t+1}(k_{t+1}), d_t)$. The effect on the saving function in Fig. 21.1 is obtained by differentiating the accumulation equation with respect to k_{t+1} and d_t, holding k_t constant. Recall, also, that $r_{t+1} = f'(k_{t+1})$.

$$(1+n)\mathrm{d}k_{t+1} = s_r f''\mathrm{d}k_{t+1} + \frac{\partial s_t}{\partial d_t}\mathrm{d}d_t \quad (21.29)$$

or

$$\frac{dk_{t+1}}{dd_t} = \frac{\frac{\partial s}{\partial d_t}}{1 + n - s_r f''(k_{t+1})} \tag{21.30}$$

The numerator is negative. Assuming $s_r > 0$ (the substitution effect of r on s dominates the income effect), as it must be for the dynamics of the economy to be reasonable, then the dominator is positive. Therefore, $\frac{dk_{t+1}}{dd_t} < 0$. The saving function shifts down at every k_t as illustrated in Fig. 21.2. The higher saving function s^0 is the one in the baseline economy without social security, and the lower saving function s^1 is the one with an unfunded social security system. The steady state level of capital decreases from k^* to k^{**}. The economy grows more slowly from an initial k_0, and moves farther away from the Golden Rule k at which consumption per person is maximized.[6]

The analysis captures only the partial equilibrium effects of an unfunded social security system on saving. With a lower k, labor is less productive and w_t falls as well. The decrease in wage income lowers saving even more. In addition, the interest rate rises as saving decreases, which also reduces saving if the substitution effect dominates. Therefore, the general equilibrium effects of introducing an unfunded social security system support the decrease in saving.

U.S. Social Security and Saving

The decrease in saving that would result from an unfunded social security system in the simple two-period OLG model would appear to apply to the United States. Alan Auerbach and Larry Kotlikoff developed a multiperiod OLG model that is far more complex than our simple two-period model but has the same general assumptions and structure, calibrated it to the U.S. economy, and used it to evaluate a number of different fiscal policies. One of the policies was the introduction of an unfunded social security pension system with the approximate characteristics of the U.S. system. Their model generates a substantial decrease in saving and investment over time, such that the capital stock, income, and consumption per capita are much lower in the new steady state. The loss in consumption per capita is the same that would occur if individuals gave up 6.9% of their lifetime resources, a huge loss (Auerbach and Kotlikoff, 1987).

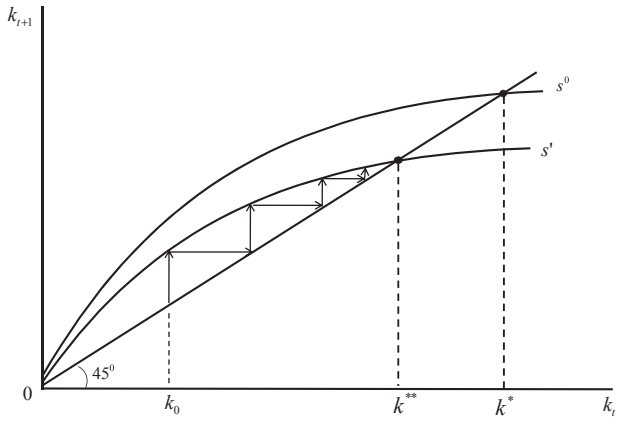

FIGURE 21.2

Given potential losses of this magnitude predicted by standard OLG macro models, economists have had a natural interest in the empirical question: Does the U.S. Social Security System reduce private saving and, if so, by how much? Surely the standard OLG macro model cannot be the last word on saving behavior. It assumes everyone is completely rational and saves according to the life-cycle hypothesis. There is plenty of evidence to suggest that other kinds of saving behavior exist, such as shorter-term precautionary saving. And many people do not save at all. In addition, our standard model ignores the bequest motive for saving. If the elderly care about the welfare of their children, they might increase their saving for bequests following the institution of an unfunded social security system to offset the redistribution from the young to the (initial) elderly (as described in the next section). This bequest behavior is the basis of the so-called Ricardian equivalence theorem, most closely associated with Robert Barro, which says that an unfunded social security system has no effect on the economy. These different kinds of behavior make it difficult to test the effect of anything on saving, which may explain why estimates of the effect of the U.S. Social Security System on the U.S. personal saving are so varied. Nonetheless, John Gruber reviewed the literature and concluded that the consensus estimate is a 30–40% reduction in private saving, a fairly hefty reduction (Gruber, 2005). If the consensus estimate is even close to accurate, the Auerbach/Kotlikoff estimate of the deleterious long-run effects, the U.S. Social Security System might be reasonably accurate.

Saving with a Defined Contribution Social Security System

The depressing effect of an unfunded social security system on saving has led to calls for reform in the United States, in particular to replace the current unfunded defined benefits

6. If the economy were originally dynamically inefficient, with k^* above the Golden Rule capital stock, then an unfunded social security system increases consumption per capita by reducing saving. Also, saving would decrease by more than the social security payment d_t. As noted above, however, the U.S. economy is undoubtedly dynamically efficient.

system with a defined contribution system. Under a defined contribution program, individuals would be forced to contribute a certain amount each year into a personal account to provide income for their retirement. Alternatively, they could pay a tax to the federal government as now and the government would invest the funds for them. The returns to the government investments would be earmarked to each individual's tax contribution and then made available to them upon retirement. The government may or may not insist that the individuals purchase an annuity with their accumulated funds upon retirement.

Having the government invest the funds is generally opposed in the United States because of the fear that the government would exert too much influence on private sector financial markets. This is especially true of government investment in common stocks, which would give the government an ownership position in businesses. Therefore, the preferred alternative in the United States is to have individuals set aside a mandatory amount each year for their retirement, with their personal accounts managed by private investment firms, who would offer individuals a variety of choices. In effect, each individual is responsible for overseeing his or her own retirement fund. The funds would be available upon retirement, and the government may or may not insist that they be used to purchase an annuity. This type of defined contribution plan was proposed by the George W. Bush administration and referred to as privatizing Social Security. It was rejected by Congress in favor of the current unfunded system. A number of countries, however—Argentina, Australia, Chile, Mexico—have "privatized" their social security systems.

The advantage of a defined contribution system is easily demonstrated in our baseline model. Consider, again, Eqn (21.25), the consumer's FOC expressed in terms of saving without a social security system:

$$U'_1(w_t - s_t) = (1+\partial)^{-1}(1+r_{t+1})U'_2(s_t(1+r_{t+1})) \tag{21.31}$$

Under a defined contribution plan, the individual places an amount d_t in an investment account when young, with the funds invested at the market rate of interest r_{t+1}. The investment grows to $d_t(1+r_{t+1})$ when old. Therefore, FOC (21.31) becomes

$$U'_1(w_t - s_t - d_t) = (1+\partial)^{-1}(1+r_{t+1})U'_2(s_t(1+r_{t+1}) + (1+r_{1+t})d_t) \tag{21.32}$$

or

$$U'_1(w_t - (s_t + d_t)) = (1+\partial)^{-1}(1+r_{t+1}) U'_2((1+r_{t+1})(s_t + d_t)) \tag{21.33}$$

The accumulation equation in terms of saving is

$$(1+n)k_{t+1} = s_t + d_t \tag{21.34}$$

Inspection of conditions (21.33) and (21.34) indicates that the defined contribution plan has no effect on the economy. If $(1+n)k_{t+1}$ is an individual's desired level of saving in the absence of a mandatory social security system, it remains the desired level of saving under a defined contribution system. With the contributions d_t invested at the market rate of interest, consumers are indifferent between the two forms of saving for their retirement. Hence they will reduce their private saving s_t dollar-for-dollar with the mandatory contribution d_t. The only caveat is that d_t must be less than $(1+n)k_{t+1}$, which may not be the case for some individuals, especially those with low incomes. If not, then total saving in the economy would rise, with some of it being forced saving. In general, though, total saving in the economy is expected to be much higher under a defined contribution plan than with an unfunded system. We will assume that saving is unaffected in our model with identical individuals, that the mandatory contribution is less than the individuals' desired saving without social security.

The Intergeneration Redistributional Effects of Social Security

A defined contribution system has an additional advantage over an unfunded system in the minds of many: It avoids the intergenerational redistributions that occur with an unfunded system, redistributions that can be quite large. These redistributions are seen as inappropriate in a program that is essentially a form of social insurance.[7]

Intergenerational redistribution is inevitable when an unfunded social security system is instituted. The initial elderly generation receives a windfall gain from the initial younger generation, in return for which they have paid nothing. The initial young generation and all succeeding generations lose because they are forced to make contributions on which they receive a rate of return equal to the rate of growth of the economy. If, instead, they could have invested these contributions on their own, they would have received the market rate of return. Assuming that the economy is dynamically efficient, the market rate of return exceeds the rate of growth of the economy, and thus they lose the difference in these rates of return on their contributions.

Under the assumptions of the baseline model in the previous section, an unfunded social security system is a

7. The analysis of this section follows closely the analysis of Feldstein and Liebman in Feldstein and Liebman, pp. 2257–2265 and 2297–2304. Feldstein is a leading advocate of switching the U.S. Social Security System to a defined contribution plan.

pure intergenerational transfer scheme. To see this, assume as above that n equals the rate of growth of the economy, in this section interpreted as the sum of the rate of growth in the population and the rate of productivity growth. r equals the market rate of return, equal to the marginal product of capital. There are no taxes on capital income, so that r also equals the rate of return to individuals' saving. The labor supply for each period is fixed and wage income grows at rate n over time. The young workers in each period contribute d_t to the social security system, which is immediately paid out to the older generation. Think of d_t as a payroll tax at a single rate on each individual's labor income.

The first generation of elderly at t_0 receives a windfall gain of d_0, the contribution of the first generation young. When the first generation of young retire, they receive $(1 + n)d_0$ from the next generation of young. Had they been able to invest their contribution at the market rate, they would have received a return of $(1 + r)d_0$. Therefore, they lose $(r - n)d_0$ on their contribution, which they would likely discount at the market rate $(1 + r)$. The present value of their loss is $(r - n)d_0/(1 + r)$. This can be thought of equivalently as a loss of income or of consumption when elderly. Every succeeding generation suffers the same loss on its contributions as the initial young do. Noting that the contributions grow at rate n each year, and discounting the losses of all generations back to present value at t_0 at the discount rate r generates an aggregate present value of loss equal to

$$\mathrm{PV}_{\mathrm{losses}} = d_0\left(\frac{r-n}{1+r}\right)\sum_{t=0}^{\infty}\frac{(1+n)^t}{(1+r)^t} \tag{21.35}$$

Given that $n < r$, in the limit $\sum_{t=0}^{\infty}\frac{(1+n)^t}{(1+r)^t} = \frac{1}{1-\frac{(1+n)}{(1+r)}} = \frac{(1+r)}{(r-n)}$. Therefore, $\mathrm{PV}_{\mathrm{losses}} = d_0$, the windfall gain to the initial elderly. The unfunded social security system is a pure transfer across generations.

The pure transfer result depends on three unrealistic assumptions, however, that: (1) The discount factor applied to the stream of losses of time is the marginal product of capital; (2) There are no capital income taxes; and (3) The labor supply is constant, such that a payroll tax or other forced contribution from wage income generates no deadweight loss. Relaxing any of these in our baseline model generates $\mathrm{PV}_{\mathrm{losses}}$ greater than the windfall gain to the initial elderly generation. Let us briefly consider each of these.

The Discount Factor

In a social welfare maximizing framework, it is natural to discount the stream of consumption losses at the social rate of time preference, ρ, above. Now,

$$\mathrm{PV}_{\mathrm{losses}} = d_0\left(\frac{r-n}{1+r}\right)\sum_{t=0}^{\infty}\frac{(1+n)^t}{(1+\rho)^t} \tag{21.36}$$

Assuming a high social rate of time preference, such that $\rho > n$, and the present value of the losses of future generations decline over time, then in the limit $\sum_{t=0}^{\infty}\frac{(1+n)^t}{(1+\rho)^t} = \frac{(1+\rho)}{(\rho-n)}$. Therefore,

$$\mathrm{PV}_{\mathrm{losses}} = d_0\left(\frac{r-n}{1+r}\right)\frac{(1+\rho)}{(\rho-n)} > d_0 \tag{21.37}$$

since the discount rate $\rho < r$, the discount rate in Eqn (21.35). The unfunded system generates a loss in social welfare in the process of redistributing across the generations. If $\rho < n$, which is highly likely, then the $\mathrm{PV}_{\mathrm{losses}}$ grows exponentially with each succeeding generation. These results are not surprising, since it takes a very high discount rate, r = marginal product of capital, to equate the $\mathrm{PV}_{\mathrm{loss}}$ to the initial windfall gain.

Capital Income Taxes

When income from capital is taxed, individuals receive the net-of-tax return, r_n, rather than a return equal to the marginal product of capital. Hence they will discount future income at r_n, not r. To trace the $\mathrm{PV}_{\mathrm{losses}}$, we have to distinguish between the decrease in consumption and saving when young, caused by their contribution. The first generation of young lose $d_0(1 - s)$ of consumption and $d_0 s$ of saving. The saving would have generated income of $(1 + r)s$, a rate equal to the marginal product of capital, if invested in the market, some of which would have gone to the government in taxes. Assume that it would have been returned lump sum to the individual so that all income in the economy is accounted for. Instead the individual receives $(1 + n)d_0$ in the unfunded plan. The income received on the saving next period is discounted at $(1 + r_n)$, the net-of-tax return. Therefore, the $\mathrm{PV}_{\mathrm{losses}}$ of consumption from the contribution is[8]

$$\mathrm{PV}_{\mathrm{losses}} = (1-s)d_0 + (s(1+r)d_0 - (1+n)d_0)(1+r_n)^{-1} \tag{21.38}$$

Multiplying and dividing the first term by $(1 + r_n)$ and collecting terms yields

$$\mathrm{PV}_{\mathrm{losses}} = d_0(1+r_n)^{-1}\left[(r_n - n) + (r - r_n)s\right] \tag{21.39}$$

8. Note that if $r = r_n$, $\mathrm{PV}_{\mathrm{losses}} = d_0\,(r-n)/(1 + r)$ as above. The split of d_0 between saving and consumption does not matter because any reduction of saving grows at rate r and is discounted at rate r. Therefore, the loss of consumption is d_0 regardless of the value of s.

For the generation that is young at time t,

$$\mathrm{PV}_{\text{losses}} = d_0(1+r_n)^{-1}\left[(r_n-n)+(r-r_n)s\right](1+n)^t \tag{21.40}$$

because d_0 grows over time at rate n. If the consumption losses in each generation are discounted by the social rate of time preference ρ, then the aggregate discounted stream of losses is

$$\mathrm{PV}_{\text{losses}} = d_0(1+r_n)^{-1}\left[(r_n-n)+(r-r_n)s\right]\sum_{t=0}^{\infty}\frac{(1+n)^t}{(1+\rho)^t} \tag{21.41}$$

In the limit

$$\mathrm{PV}_{\text{losses}} = d_0(1+r_n)^{-1}\left[(r_n-n)+(r-r_n)s\right]\frac{(1+\rho)}{(\rho-n)} > d_0 \tag{21.42}$$

for reasonable values of r, and n and a high social rate of time preference ρ ($>n$).[9]

Variable Labor Supply and Deadweight Loss

If labor supply is variable, then a tax on labor income (or any mandated contribution from labor income) may generate an additional cost to all but the initial elderly in the form of a deadweight loss. As we saw in Chapter 13, the marginal deadweight loss per dollar of revenue collected is the product of the marginal tax rate and the compensated elasticity of the supply of labor with respect to the wage: marginal loss $= tE^C_{l,w}$.[10] An important issue, however, is how people view the tax (contribution). Consider, first, a defined contribution plan. As we have seen, the tax (contributions) provides income for retirement and earns the market rate of return. It is indistinguishable from private saving in this regard. Hence, it would be viewed as a benefits-received tax, and there is no deadweight loss associated with benefits-received taxes.[11]

The payroll tax (contribution) in an unfunded system might also be viewed as a benefits-received tax, since payment of the tax comes with an implicit promise that future generations will provide retirement benefits to the current workers. But the analogy to a defined contribution plan is not exact, for a number of reasons. First, the unfunded system implies a loss of income on the contribution equal to the difference in the market rate of return and the growth of the economy, and the loss of income could affect the supply of labor. The effective tax rate is the payroll tax rate times $(r-n)/r$, the proportional loss in income on the taxes paid. Second, under the U.S. Social Security System, only the highest 35 years of earnings are used to calculate the retirement benefits. Young workers may understand that their payroll taxes in their first years of work are likely to have no effect on their eventual retirement benefits. If so, then they would view the tax payments as a general tax that could generate a deadweight loss. Finally, spouses who are second earners in a family can receive a benefit of no more than half of the benefit of the higher earner if the higher earner receives the maximum benefit. For them, much of their payroll tax payments will essentially be just another general tax on their labor earnings, and one that could generate a deadweight loss. Martin Feldstein believes that all these factors are important sources of deadweight loss, enough so that he estimates that the payroll tax generates a marginal deadweight loss of 50% per dollar of revenue collected (Feldstein, 2005). It is to be noted that 50% is on the high end of deadweight loss estimates of the payroll tax, but there is little doubt that the payroll tax is not just a benefits-received tax.

In summary, the redistributions to the initial generation in an unfunded social security system are likely to lead to a loss of efficiency and social welfare in actual economies. They are not neutral.

Social Security Reform: Switching to a Defined Contribution Plan

Many economists supported the call by President George W. Bush to privatize the U.S. Social Security System by turning it into a defined contribution plan. They saw this reform as a way of increasing the U.S. saving and investment and of avoiding the intergenerational redistributions inherent in the current unfunded system, redistributions that almost certainly reduce social welfare. But the transition in switching from an unfunded to a fully funded defined contribution plan is not easily managed, which may have been the tipping point for many Congressmen in voting to maintain the current system.

The transition difficulties derive from having to offer the benefits already earned by the workers and retirees in the current unfunded system while phasing in the defined contribution plan. The two-period OLG model offers a simple framework for analyzing the essence of the transition. It allows us to assume that the young workers have not yet made any contributions to the current unfunded system—the switch occurs at the beginning of their working years. Only the benefits earned by the current retirees need to be covered

9. If $r = r_n$, then $\mathrm{PV}_{\text{losses}}$ is that given in Eqn (21.37).
10. We are assuming here for simplicity that no other markets are taxed or distorted, so that no cross-price elasticities are included in the loss.
11. This assumes that the tax (contribution) is less than the individual's desired saving without the public defined contribution plan. An individual surely suffers a loss of utility if the tax exceeds his or her desired saving, as long as the individual is rational about providing for retirement.

during the transition. Once the initial retirees die at the end of the first period, the transition is complete.

Martin Feldstein and Jeffrey Liebman produced an example of a transition that makes use of recognition bonds, a method used by Chile when it switched to a defined contribution plan (Feldstein and Liebman, pp. 2297–2302). In the transition period, t_1, the current young workers stop paying into the unfunded system. Instead they place the tax payments d_1 they would have made into their own personal retirement accounts, to be invested at the market rate of interest r, assumed constant over time. The current retirees receive recognition bonds equal to d_1 that they can use to finance their consumption in retirement, so-named because the bonds are in recognition of the obligation owed to the retirees under the replaced unfunded system. The bonds are issued in perpetuity, and pay interest rd_1 each year at the market rate. The advantage of issuing recognition bonds for the transition is that it spreads the burden of paying for the obligations to the current elderly over all generations. In Feldstein/Liebman's example, the young workers of each generation contribute the same amount to their personal accounts as they would have contributed under the replaced unfunded system. Hence their contributions grow each year by $(1 + n)$, interpreted here as the rate of growth of the economy. The change in aggregate consumption each year is as follows.

accounts. The gains eventually come because the initial decreases in consumption increase saving and the stock of capital. An increase in national saving has to increase capital, and thus income and consumption, if the economy is dynamically efficient.

Adding up the net gains over time from switching to a defined contribution plan, discounted at the social rate of time preference ρ, yields

$$\text{PV}_{\text{consumption}} = \sum_{t=1}^{\infty} d_1(r-n)\frac{(1+n)^{t-1}}{(1+\rho)^t} \tag{21.43}$$

the first two lines in the Table.

$$\text{PV}_{\text{debt burden}} = -\sum_{t=1}^{\infty} rd_1\frac{1}{(1+\rho)^t} \tag{21.44}$$

third line in the Table.

$$\text{Net PV} = \sum_{t=1}^{\infty} d_1(r-n)\frac{(1+n)^{t-1}}{(1+\rho)^t} - \sum_{t=1}^{\infty} rd_1\frac{1}{(1+\rho)^t} \tag{21.45}$$

In the limit,

$$\text{Net PV} = \left[\frac{(r-n)}{(\rho-n)} - \frac{r}{\rho}\right] d_1 > 0 \tag{21.46}$$

	t_1 (Initial Transition Year)	t_2	t_3	t_4
Retirees	d_1	$d_1(1 + r)$	$d_1(1 + n)(1 + r)$	$d_1(1 + n)^2(1 + r)$
Workers	$-d_1$	$-d_1(1 + n)$	$-d_1(1 + n)^2$	$-d_1(1 + n)^3$
Debt service	0	$-rd_1$	$-rd_1$	$-rd_1$
Change in aggregate consumption	0	$-nd_1$	$d_1((1 + n)(r - n) - r)$	$d_1((1 + n)^2(r - n) - r)$

A worker at time t is a retiree at time $t + 1$ and earns $(1 + r)$ on the money placed in his/her personal account at time t. The changes in aggregate consumption every year after the transition are the net result of two opposing effects. The negative effect is the initial lowering of the capital stock by the amount of the recognition bonds, which is borne by the perpetual burden of paying interest on the debt. The positive effect is that workers now receive the market rate r rather than the rate n on their retirement contributions. At first the negative effect of paying for the obligations of the initial retirees dominates, and aggregate consumption falls. This transition burden is represented in the example by the loss of consumption of nd_1 in period t_2. But then the positive effect kicks in, and eventually aggregate consumption will increase. It increases by ever-larger amounts thereafter as the economy grows and workers increase their contributions to their personal

under the reasonable assumptions that $r > n$ (the economy is dynamically efficient), $r > \rho$ (the market rate of interest exceeds the social rate of time preference), and $n > 0$ (the economy is growing). The limit also assumes that $\rho > n$, a high social rate of time preference.

Another advantage of switching to a defined contribution plan is a reduction in the deadweight loss associated with the taxes used to finance the contributions given that the supply of labor is variable. As noted above, the taxes paid into a defined contribution plan are truly benefits-received tax because they earn the market rate of return. They cannot be a source of deadweight loss. The only qualifier is that a government may use an increase in taxes rather than recognition bonds to finance the existing obligations in the unfunded system. This was the plan under the George W. Bush proposal to privatize U.S. Social Security. The temporary increase in taxes would increase the

deadweight loss of the payroll tax during the transition but eventually the taxes will be lowered and will be lower than in the unfunded system to finance the same stream of benefits. In the Feldstein/Liebman example above, financing the initial retiree obligation with an increase in taxes on the next few generations increases the burden far more than the recognition bonds do and would likely lead to losses for a number of periods. But the long-run gains would eventually appear because of the ability to invest contributions at the higher market rate, and when they do, the deadweight losses from the payroll tax would decrease. Overall, the decreases in deadweight loss in the long run more than offset the result in the increases in dead-weight loss in the short run.

The Legacy Debt in the United States

The need to meet the obligations in the unfunded system during the transition turns out to be a huge problem for the United States. When Social Security was initiated in 1935, the payroll tax rate was only 2% and relatively few workers were covered. From 1935 to 1983, there were both increases in the payroll tax rates to 9.55% on the pensions by 1983 and a huge expansion in the number of workers covered. The increased revenues that resulted from these to factors were mostly paid out to retirees to increase their benefits. It was not until 1983 that the government attempted to place Social Security on a fully funded basis through the retirement of the baby boomers, by phasing in future legislated tax increases more quickly and increasing the retirement age for receiving full benefits from 65 to 67 years over a number of years. Consequently, anyone retiring before 1983 received much larger benefits than they ever contributed; the average annual returns on their contributions were above even the marginal product of capital, the before-tax rate of return to capital. It was as if the windfall gain to the initial retirees in the simple OLG model occurred for a number of generations. These windfall gains ended post-1983, but Social Security was far from being on a fully funded basis.

By 2004, when privatization was being debated, there were enormous unfunded obligations remaining to retirees and workers who had contributed to Social Security. Peter Diamond and Peter Orszag calculated that the pre-1983 retirees placed what they called a legacy debt burden on the system of $11.6 trillion, equal to the amount they should have been contributing to the Social Security Trust Fund over time plus interest on those missing contributions, were the system fully funded from its inception (Diamond 2004, p.16). The huge legacy debt makes switching to a defined contribution plan extremely difficult. The transition is possible, however. Martin Feldstein and Andrew Samwick worked out a feasible, but long, transition in which individuals contribute to both the paygo system and a defined contribution plan. Initially the combined contribution is only 1.6% points above the current payroll tax rate. It takes 25 years for the combined rate to drop below the current payroll tax rate, and the paygo system is phased out completely in 75 years. The defined contribution rate after the phase out is only 3.25%. Their scheme used the 75-year demographic and economic projections of the Social Security Trustees at the time.[12]

The legacy debt from the earlier generations is a problem even for the existing system. Current estimates are that the assets of the Trust Fund will be completely depleted by 2033, well before all the baby boomers have died. The 1983 reforms turned out to be too optimistic in being able to build up the Trust Fund sufficiently to finance the baby boomers' retirement. The depletion of the Trust Fund bothers many people who want to maintain the current defined benefit system; they would like it to be fully funded going forward. But the legacy debt makes it politically infeasible to raise taxes by enough to fully fund the Trust Fund going forward. The most anyone is seriously proposing is to increase taxes by enough so that the Trust Fund becomes sustainably solvent, meaning that the ratio of Fund assets to annual expenditures is either constant or growing. Only a relatively small increase in payroll tax revenues, equal to about 2.6% of GDP, is necessary to achieve sustainable solvency (Diamond 2004, p.1).

CONCLUDING OBSERVATIONS

The standard OLG model makes a strong case for switching to a defined contribution pension system over the unfunded defined benefits system that the United States has chosen. The model predicts that this would increase the rate of saving, and therefore social welfare in the long run given that the U.S. economy is dynamically efficient. It would also avoid the intergenerational redistributions of the current system, redistributions that are social welfare reducing.

The conclusion to switch to a defined contribution is hardly surprising given that the OLG model, with its rational LCH savers, really has no use for any type of public pension. If one were introduced, it would make sense to institute a defined contribution plan that has no effect on anything rather than an unfunded defined benefits plan with its harmful effect on national saving and the welfare-reducing intergenerational redistributions. Nonetheless, the model is suggestive of some of the key practical issues that might favor one plan over the other.

12. Feldstein and Samwick (1998 pp.215–260). The outline of the plan presented in the text is described in Feldstein and Liebman pp.2035–2036.

Peter Diamond and Martin Feldstein debated the merits of the two plans in their Presidential Addresses to the American Economic Association that appeared 1 year apart in the *American Economic Review* in 2004 and 2005, when the issue of privatizing the U.S. Social Security System was being debated (Diamond, 2004; Feldstein, 2005). Diamond favored the status quo, whereas Feldstein supported a switch to a defined contribution plan. They differed on four main points:

First, Diamond is persuaded by the social insurance arguments in favor of the current unfunded system. He is willing to take a paternalistic approach to individuals' saving decisions, arguing that far too many people would be incapable of the long-run planning necessary to adequately fund their retirement. A particular problem is that they misunderstand the advantages of annuities, which leads him to favor the defined benefits annuity provided by the U.S. system. A defined contribution plan could force individuals to annuitize their assets when they retire, but forcing people to do this is unlikely to be persuasive to those who favor a defined contribution plan. Feldstein distrusts paternalistic arguments. He believes it is better to assume people are rational and can plan adequately for their retirement, and thus it makes sense to let them earn the higher returns available under a defined contribution plan. As noted above, he also believes the switch could be handled without being too much of a burden.

Second, Diamond likes the intragenerational redistribution built into the Social Security benefit formula and is not bothered by the intergenerational redistributions to the older generations, given that the younger generations are increasingly better off when the economy is growing. Feldstein believes that a public pension plan should not be redistributive, at least not across generations.

Third, Diamond believes that the moral hazard problems with the current system are minimal. The direct moral hazard issue is the incentive to retire early, given that workers can begin to collect benefits at 62 years of age. The indirect moral hazard issue is the deadweight loss from the payroll tax, which Diamond sees as unimportant. He believes that most beneficiaries view the payroll tax as benefits-received tax. As noted earlier, Feldstein disagrees completely about the deadweight loss of the payroll tax, believing that the marginal deadweight loss of the tax is on the order of 0.50 per dollar of tax revenue.

Fourth, the current unfunded system is susceptible to political risk, that the administration and Congress will reduce the benefits that people had expected to receive for their contributions. A defined contribution plan is susceptible to the standard market risks associated with investing in stocks, bonds, and other kinds of securities. The market risks are irrelevant to sophisticated investors who hold diversified portfolios from risky common stocks to virtually riskless U.S. Treasury securities. For them, the higher average market return available on common stocks relative to Treasury securities just compensates for the extra risk they entail. It is the smaller savers and those who would otherwise not save at all who are most susceptible to market risk. Diamond believes that the political risk to the current system is quite small, and that Social Security is almost untouchable politically. Feldstein is unconcerned about the market risk of a defined contribution plan. He thinks the odds are extremely small that the market return would not exceed the rate of growth of the economy over the long run.[13]

Finally, Feldstein believes that the switch to a defined contribution system would not be too costly if phased in slowly over time while the unfunded system is slowly being phased out. Diamond is more concerned about the legacy debt of the pre-1983 generations, which makes it politically difficult to make the Social Security Trust Fund even sustainably solvent. As it happened, the call for privatization fell flat politically and came nowhere close to passing. This is hardly surprising since U.S. citizens, and Congress, at the time were in no mood to try something new that would place an additional burden on workers for a number of years.

REFERENCES

Auerbach, A., Kotlikoff, L., 1987. Dynamic Fiscal Policy. Cambridge University Press, New York, p. 153.

Blanchard, O., Fisher, S., 1989. Lectures on Macroeconomics. The MIT Press, Cambridge, Massachusetts.

Diamond, P., December 1965. National debt in a neoclassical growth model. American Economic Review 55 (5), 1126–1150.

Diamond, P., March 2004a. Social security. American Economic Review 94 (1), 1–24.

Diamond P., 2004. Social Security. Op. Cit., p. 16.

Diamond P., 2004. Social Security. Op. Cit., p. 1.

Diamond P. Social security. Op. Cit., and Feldstein M. Rethinking social security. Op. Cit.

Feldstein M., Liebman J. Social security. Op. Cit. Section 3, pp. 2257–2265 and Section 7, pp. 2297–2304.

Feldstein, M., March 2005. Rethinking social security. American Economic Review 95 (No. 1), 9.

Feldstein, M., Liebman, J., 2002. Social security. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier Science B.V, p. 2256.

Feldstein M., Liebman J. Social security. Op. Cit., pp. 2035–2036.

Feldstein M., Liebman J. Social Security. Op. Cit., Sect. 7.1.3, pp. 2297–2302.

13. Ironically, had privatizing been legislated, many people would no doubt have seen it as a catastrophic move given the financial crisis of 2008, even though that is too short a time frame to judge a defined contribution plan.

Feldstein, M., Samwick, A., 1998. The transition path in privatizing social security. In: Feldstein, M. (Ed.), Privatizing Social Security. University of Chicago Press, Chicago, pp. 215–260.

Gruber, J., 2005. Public Finance and Public Policy. Worth Publishers, New York, p. 344.

Munnell, A., Sunden, A., 2006. 401K Plans and Still Coming Up Short. An Issue in Brief, Number 43. Center for Retirement Research, Boston College. Table 1, p. 2.

2012 Annual Report of the Trustees of the Federal Old Age and Survivors Insurance and Federal Disability Insurance Trust Funds, www.ssa.gov/oact/tr/2012/index.html.

Samuelson, P., December 1958. An exact consumption loan model of interest with or without the contrivance of money. Journal of Political Economy 66 (6), 467–482.

Chapter 22

Externalities in a Second-Best Environment

Chapter Outline

We saw in Chapters 6–8 that first-best models of externalities dichotomize in two important respects for policy purposes. One is that the government can pursue appropriate tax or expenditure policies to restore pareto optimally in the presence of externalities without regard to distributional considerations. All distributional issues are embodied in the interpersonal equity conditions, which can be satisfied by an appropriate set of lump-sum taxes and transfers. The other is that externalities arising within a subset of all goods and factor markets can be corrected independently of behavior in the other markets, in the sense that the perfectly competitive allocations in these markets remain pareto optimal. These two properties greatly facilitate policy design when correcting for externalities.

Unfortunately, neither of these dichotomies holds in a second-best environment. As a consequence, even the simplest externalities may require highly complex forms of government intervention, so complex in fact that it is entirely implausible to expect governments to achieve them. To illustrate this fundamental point, we will consider the example of providing a Samuelsonian nonexclusive consumption good in a many-person economy made second best because the government does not have the ability to tax and transfer lump sum to achieve the first-best interpersonal equity conditions.

THE SECOND-BEST ALLOCATION OF SAMUELSONIAN NONEXCLUSIVE GOODS

The first-best analysis of a Sameulsonian nonexclusive public good yielded three specific policy prescriptions:

1. The government should provide the good such that $\sum_{h=1}^{H} \text{MRS}^h = \text{MRT}$. The government has to provide the good because the incentive to free ride prevents the market system from allocating nonexclusive goods.
2. If the government happens to select the quantity that satisfies the optimal decision rule, then it can finance the good with any lump-sum tax. The lump-sum tax keeps the economy on the first-best utility–possibilities frontier. Any unwanted distributional consequences of the tax are overcome by the lump-sum taxes and transfers that satisfy the first-best interpersonal equity conditions for a social welfare maximum.
3. The competitive market economy can be counted on to generate the pareto-optimal allocations of all the purely private goods and factors.

None of these prescriptions applies in a second-best environment in general, although the ways in which the

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00022-4

first-best optimal decision rules change depend upon the nature of the additional constraints placed on the system. This is always true in second-best analysis. A natural way to pose a second-best problem is to let the government freely choose the quantity of the nonexclusive good but constrain it to finance the good with distorting unit commodity taxes. This implicitly precludes lump-sum redistributions to satisfy the first-best interpersonal equity conditions by equalizing marginal social utilities of income, because if lump-sum taxes could be used for distributional purposes, they should also be available to finance the public good. Otherwise, assume that the economy is perfectly competitive with all other goods (factors) being purely private. In other words, the need to use distorting taxes is the only constraint that makes the analysis second best.

Given this particular second-best environment, there are two compelling policy questions to be asked:

1. How does the required distorting taxation affect the optimal decision rule for providing the public good?
2. How does the presence of the public good affect the optimal tax rules when revenue is raised for its own sake?

These questions can be addressed with a general equilibrium model that is a straightforward extension of the many-person model used in Chapter 14 to analyze optimal commodity taxation under general technology.

Preferences and Social Welfare

Let e stand for the nonexclusive good, defined in units such that its price equals 1. Since the government is selecting the quantity of e, consumers treat e as a parameter even though e enters their utility functions. Therefore, each individual solves the following utility maximization problem:

$$\max_{(X_{hi})} U^h(X_{hi}; e)$$
$$\text{s.t.} \sum_{i=1}^{N} q_i X_{hi} = \bar{I}^h$$

where

q_i = the consumer price of good (factor) i, $i = 1, \ldots, N$.
X_{hi} = good (factor) i consumed (supplied) by person h, $i = 1, \ldots, N$.
$h = 1, \ldots, H$.
$\bar{I}^h$ = the fixed amount of lump-sum income for person h, which the government cannot change through lump-sum redistributions.

The consumer's maximization problem leads to demand (factor supply) functions of the form:

$$X_{hi} = X_{hi}\left(\vec{\mathbf{q}}; \bar{I}^h; e\right) \quad i = 1, \ldots, N; \quad h = 1, \ldots, H \tag{22.1}$$

and indirect utility functions:

$$U^h\left[X_{hi}\left(\vec{\mathbf{q}}; \bar{I}^h; e\right)\right] = V^h\left(\vec{\mathbf{q}}; \bar{I}^h; e\right) \quad h = 1, \ldots, H \tag{22.2}$$

Social welfare, then, is

$$W^*\left[U^h\left(X_{hi}\left(\vec{\mathbf{q}}; \bar{I}^h; e\right)\right)\right] = W\left[V^h\left(\vec{\mathbf{q}}; \bar{I}^h; e\right)\right] \tag{22.3}$$

where $W(\ )$ is the Bergson−Samuelson individualistic social welfare function.

Production and Market Clearance

e must also enter the aggregate production function F because it uses real resources.[1] Therefore, write the aggregate production function implicitly as

$$F(X_i; e) = 0 \quad i = 1, \ldots, N \tag{22.4}$$

where X_i is the aggregate demand (supply) for good (factor) i. Assume that $F(\ )$ exhibits constant returns to scale so that there are no pure profits in the economy. Finally, incorporate market clearance directly into the aggregate production function:

$$F\left[\sum_{h=1}^{H} X_{hi}\left(\vec{\mathbf{q}}; \bar{I}^h; e\right); e\right] = 0$$

Social Welfare Maximization

Society's problem, then, is[2]

$$\max_{(q_i; e)} W\left[V^h\left(\vec{\mathbf{q}}; \bar{I}^h; e\right)\right]$$
$$\text{s.t. } F\left[\sum_{h=1}^{H} X_{hi}\left(\vec{\mathbf{q}}; \bar{I}^h; e\right); e\right] = 0$$

with the corresponding Lagrangian equation:

$$\max_{(q_i; e)} L = W\left[V^h\left(\vec{\mathbf{q}}; \bar{I}^h; e\right)\right] - \lambda \cdot F\left[\sum_{h=1}^{H} X_{hi}\left(\vec{\mathbf{q}}; \bar{I}^h; e\right); e\right]$$

1. e could be privately produced like missiles and military aircraft.
2. Recall that maximizing W with respect to $\vec{\mathbf{q}}$ is equivalent to maximizing W with respect to $\vec{\mathbf{t}}$, with $\vec{\mathbf{q}} = \vec{\mathbf{t}} + \vec{\mathbf{p}}$ and the market clearance equations establishing the relationships among $\vec{\mathbf{q}}$, $\vec{\mathbf{t}}$, and $\vec{\mathbf{p}}$ in equilibrium. Also, good 1 is the untaxed numeraire, with $q_1 = p_1 = 1, t_1 = 0$. This is the model used by Peter Diamond in Diamond, 1975.

Recall from the discussion of this type of model in Chapter 14 that the government's budget constraint,

$$\sum_{h=1}^{H}\sum_{i=2}^{N} t_i X_{hi} = e$$

is implied by Walras' law under the assumptions of utility maximization, profit maximization, and market clearance in all markets.

Consider the first-order conditions with respect to the consumer prices, q_i:

$$\sum_{h=1}^{H}\frac{\partial W}{\partial V^h}\frac{\partial V^h}{\partial q_k} = \lambda\sum_{i=1}^{N} F_i\frac{\partial X_i}{\partial q_k} \quad k = 2,\ldots,N \tag{22.5}$$

Conditions (22.5) are identical to conditions (14.64); therefore, the existence of a nonexclusive good does not affect the form of the many-person optimal tax rule relative to the case in which the government simply raises revenue for its own sake. Of course, the choice of e determines the amount of revenue required, which in part determines the level of tax rates, but otherwise the optimal tax rules have the identical interpretations developed in Chapter 14. The reappearance of the optimal tax rules is very discouraging in one respect. It implies that the need to finance the public good with distorting taxes forces the government to intervene pervasively into the market economy. In general, the government must tax or subsidize *all* goods and factors (except the numeraire) to achieve the second-best optimum. The provision and financing of the public good cannot be isolated from the rest of the economy as it can in a first-best environment. It could be argued that the problem resides with the commodity taxes and not with the public good itself. Remember, though, that the optimal decision rule for the public good that we are about to develop requires that Eqn (22.5) must hold for the distorting taxes. Otherwise, the first-order condition for the public good that follows would not be the necessary condition for a social welfare optimum, in general.

With these comments in mind, consider the first-order condition with respect to e:

$$\sum_{h=1}^{H}\frac{\partial W}{\partial V^h}\frac{\partial V^h}{\partial e} = \lambda\sum_{h=1}^{H}\sum_{i=1}^{N} F_i\frac{\partial X_{hi}}{\partial e} + \lambda F_e \tag{22.6}$$

Defining F such that $\partial F/\partial X_1 = 1$, assuming profit maximization under perfect competition, and given that $p_1 \equiv q_1 \equiv 1$, the untaxed numeraire, Eqn (22.6), can be rewritten as

$$\sum_{h=1}^{H}\frac{\partial W}{\partial V^h}\frac{\partial V^h}{\partial e} = \lambda\sum_{h=1}^{H}\sum_{i=1}^{N} p_i\frac{\partial X_{hi}}{\partial e} + \lambda F_e \tag{22.7}$$

But $p_i = q_i - t_i$, $i = 1, \ldots, N$. Therefore,

$$\sum_{h=1}^{H}\frac{\partial W}{\partial V^h}\frac{\partial V^h}{\partial e} = \lambda\sum_{h=1}^{H}\sum_{i=1}^{N}(q_i - t_i)\frac{\partial X_{hi}}{\partial e} + \lambda F_e \tag{22.8}$$

Next, differentiate each consumer's budget constraint, $\sum_{i=1}^{N} q_i X_{hi}(\vec{\mathbf{q}}; \overline{I}^h; e) = \overline{I}^h$ with respect to e:

$$\sum_{i=1}^{N} q_i\frac{\partial X_{hi}}{\partial e} = 0 \quad h = 1,\ldots,H \tag{22.9}$$

Substituting Eqn (22.9) into (22.8) yields

$$\sum_{h=1}^{H}\frac{\partial W}{\partial V^h}\frac{\partial V^h}{\partial e} = -\lambda\sum_{i=1}^{N} t_i\frac{\partial X_i}{\partial e} + \lambda F_e \tag{22.10}$$

Peter Diamond proposed the following $\sum_{h=1}^{H}\text{MRS}^h = \text{MRT}$ interpretation of condition (22.10) (Diamond, 1975, p. 341). Rewrite Eqn (22.10) as

$$\sum_{h=1}^{H}\frac{\partial W}{\partial V^h}\frac{\partial V^h}{\partial e} + \lambda\sum_{i=1}^{N} t_i\frac{\partial X_i}{\partial e} = \lambda F_e \tag{22.11}$$

The right-hand side (RHS) of Eqn (22.11) measures the marginal social cost, through production, of increasing the public good. The left-hand side (LHS) is the social marginal value of increasing the public good. The first term represents the social marginal value of having each person consume an additional unit of nonexclusive e; the second term represents the social value of the increased tax revenues resulting from a marginal increase in e. Thus, Eqn (22.11) has the natural interpretation that e should be increased until its social marginal value just equals its social marginal cost. To change this to a $\sum_{h=1}^{H}\text{MRS}^h = \text{MRT}$ form, define

$$\delta^h = \frac{\partial W}{\partial V^h}\frac{\partial V^h}{\partial e} + \lambda\sum_{i=1}^{N} t_i\frac{\partial X_{hi}}{\partial e} \tag{22.12}$$

Recall that the social marginal utility of income, β^h, is a product of the marginal social welfare weight and the private marginal utility of income, or $\beta^h = (\partial W/\partial V^h)\,\alpha^h = (\partial W/\partial V^h)\,(\partial V^h/\partial I^h)$. Therefore, δ^h can be expressed as

$$\delta^h = \beta^h\left(\frac{\frac{\partial V^h}{\partial e}}{\frac{\partial V^h}{\partial I^h}}\right) + \lambda\sum_{i=1}^{N} t_i\frac{\partial X_{hi}}{\partial e} = \beta^h\text{MRS}^h_{e,X_{h1}} + \lambda\sum_{i=1}^{N} t_i\frac{\partial X_{hi}}{\partial e} \tag{22.13}$$

the social marginal value of letting person h consume an additional unit of e.[3] Substituting Eqn (22.13) into (22.11) yields

$$\sum_{h=1}^{H}\delta^h = \lambda F_e \tag{22.14}$$

3. $\text{MRS}^h_{e,X_{h1}} \equiv \frac{\frac{\partial V^h}{\partial e}}{\frac{\partial V^h}{\partial X_{h1}}} = \frac{\frac{\partial V^h}{\partial e}}{\alpha^h q_1} = \frac{\frac{\partial V^h}{\partial e}}{\frac{\partial V^h}{\partial I^h}}$, from utility maximization and $q_1 = 1$.

or

$$\sum_{h=1}^{H} \frac{\delta^h}{\lambda} = F_e \tag{22.15}$$

With F defined, such that $F_1 = 1$ and $P_1 \equiv q_1 \equiv 1$, the RHS of Eqn (22.15) is the marginal rate of transformation (MRT) between the public good and the numeraire good. To interpret the LHS, recall from Chapter 14 that if the government offers an optimal equal-value head subsidy to all individuals, λ can be interpreted as the average social marginal utility of income, equal to $\sum_{h=1}^{H} \gamma^h/H$, where $\gamma^h = \beta^h + \lambda \sum_{i=1}^{N} t_i(\partial X_{hi}/\partial I)$, the social marginal utility of giving additional income to person h. Given this interpretation of λ, the LHS of Eqn (22.15) can be interpreted as $\sum_{h=1}^{H} \text{MRS}^h$, the sum of the *social* marginal rate of substitution (MRS) between consumption of e by each individual and income (or, equivalently, the numeraire good) averaged over the population.

Relationships between First-Best and Second-Best Allocations

Diamond's interpretation of the social MRS is obviously far removed from the usual notion of a social MRS for nonexclusive goods from first-best analysis. There is no obvious quantitative relationship between the first-best and second-best decision rules for the allocation of e. Clearly, the true social MRS (the Diamond measure) could be arbitrarily larger or smaller than the first-best social MRS depending upon the choice of β^h, the social marginal utilities of income.

It has long been common wisdom that a nonoptimal income distribution requires dividing the benefits (and costs) of public projects into socially relevant components and weighting each component by the appropriate social marginal utilities of income. But notice that even if the income distribution is optimal, such that $\beta^h = \beta$, for all $h = 1, \ldots, H$, the straight summation of individual MRS^h still misrepresents the true social MRS if distorting taxes are used to finance these public projects, since the tax term $\lambda \sum_{i=1}^{N} t_i \frac{\partial X_i}{\partial e}$ remains as part of the true social MRS. This point was established formally by Anthony Atkinson and Nicholas Stern and elaborated by David Wildasin, although A. C. Pigou presented an intuitive analysis as early as 1947 (Atkinson and Stern, 1974; Pigou, 1947; Wildasin, 1984).

To isolate the effect of the tax term on the social valuation of nonexclusive goods, assume that all consumers have identical tastes and endowments, $\bar{I} = \bar{I}^1, \ldots, \bar{I}^h, \ldots, \bar{I}^H$. Further, let $\partial W/\partial V^h = 1$, for $h = 1, \ldots, H$, so that the distribution is optimal from society's point of view. Hence, $\beta^h = \beta = \alpha = \partial V^h/\partial I^h$, for $h = 1, \ldots, H$, the common private marginal utility of income. Under these assumptions, Eqn (22.15) becomes (using Eqn (22.13))

$$\frac{\alpha}{\lambda}\left(H \cdot \text{MRS}^h_{e,X_{h1}}\right) + \sum_{i=1}^{N} t_i \frac{\partial X_i}{\partial e} = F_e = \text{MRT}_{e,X_1} \tag{22.16}$$

where

$(H \cdot \text{MRS}^h_{e,X_{hi}})$ = the standard first-best interpretation of the social MRS for a nonexclusive good.

According to Eqn (22.16), the true second-best social MRS (the entire LHS of Eqn (22.16)) tends to exceed the first-best social MRS the more increasing the public good increases tax revenues through its effect on the demands (supplies) of all other goods (factors), and vice versa. Assuming that the revenues increase, this provides an additional source of marginal social value that the first-best measure misses.

Suppose, however, that all purely private demands (and factor suppliers) are independent of e ($\partial X_i/\partial e = 0$, for $i = 1, \ldots, N$), so that the revenue effect vanishes. The true social MRS still differs from the first-best measure by a factor α/λ in a world with distorting taxes. The question remains, then, whether the first-best measure under- or overstates the true measure; that is, whether $\alpha/\lambda \lessgtr 1$.

α/λ can be evaluated if we assume the government is raising tax revenue optimally. With identical consumers and $\beta^h = \beta = \alpha$, and using Roy's identity, the first-order conditions for optimal taxation, Eqn (22.5), become

$$-H\alpha X_{hk} = \lambda \sum_{h=1}^{H} \sum_{i=1}^{N} F_i \frac{\partial X_{hi}}{\partial q_k} = \lambda H \cdot \sum_{i=1}^{N} F_i \frac{\partial X_{hi}}{\partial q_k} \qquad k = 2, \ldots, N \tag{22.17}$$

Reproducing the derivation of the optimal rule in Chapter 14:

$$-\alpha X_{hk} = \lambda \sum_{i=1}^{N} F_i \frac{\partial X_{hi}}{\partial q_k} \tag{22.18}$$

$$-\alpha X_{hk} = \lambda \sum_{i=1}^{N} p_i \frac{\partial X_{hi}}{\partial q_k} \tag{22.19}$$

$$-\alpha X_{hk} = \lambda \sum_{i=1}^{N} (q_i - t_i) \frac{\partial X_{hi}}{\partial q_k} \tag{22.20}$$

$$-\alpha X_{hk} = \lambda\left(-X_{hk} - \sum_{i=1}^{N} t_i \frac{\partial X_{hi}}{\partial q_k}\right) \tag{22.21}$$

$$-\alpha X_{hk} = \lambda\left(-X_{hk} - \sum_{i=1}^{N} t_i S_{ik}^h + X_{hk} \sum_{i=1}^{N} t_i \frac{\partial X_{hi}}{\partial I}\right) \tag{22.22}$$

where

$S_{ik}^h = \left.\frac{\partial X_{hi}}{\partial q_k}\right|_{\text{compensated}}$ is the Slutsky substitution term.

Dividing both sides by $-\lambda X_{hk}$ and rearranging terms:

$$\left(\frac{\alpha}{\lambda} - 1 + \sum_{i=1}^{N} t_i \frac{\partial X_{hi}}{\partial I}\right) = \frac{\sum_{i=1}^{N} t_i S_{ik}^h}{X_{hk}} \quad k = 2, \ldots, N \tag{22.23}$$

Next, multiply the numerator and denominator of the RHS of Eqn (22.23) by t_k and sum over $k = 1, \ldots, N$ to obtain

$$N \cdot \left(\frac{\alpha}{\lambda} - 1 + \sum_{i=1}^{N} t_i \frac{\partial X_{hi}}{\partial I}\right) = \frac{\sum_{i=1}^{N} \sum_{k=1}^{N} t_i S_{ik}^h t_k}{\sum_{k=1}^{N} t_k X_{hk}} \tag{22.24}$$

As long as total tax revenue$\left(\sum_{k=1}^{N} t_k X_{hk}\right)$ is positive,[4] the RHS of Eqn (22.24) is negative because the Slutsky matrix is negative definite. Other things being equal, therefore, the RHS tends to lower the value of α/λ and thereby reduce the value of the true social MRS. Pigou identified this as the "indirect damage" of having to raise additional revenues with distorting taxes to finance increases in the public good (Pigou, 1947, p. 34). The second effect involves the term $\sum_{i=1}^{N} t_i(\partial X_{hi}/\partial I)$, which Atkinson and Stern call the "revenue effect" of distorting taxes.[5] If this term is positive (that is, if tax collections rise with increases in lump-sum income), then α/λ must be less than one and the first-best social MRS overstates the true social MRS. The tax term could well be negative, however, because factor supplies are subsidized if goods are taxed (recall that factors enter the analysis with a negative sign). If so, then α/λ may be greater than, less than, or equal to 1 despite the (negative) distortionary effect of second-best taxes. Hence, there is no way of knowing, a priori, whether the true social MRS is less than, greater than, or equal to the first-best social MRS in the presence of distorting taxes, even if: (1) the distribution of income is optimal, (2) there are no direct revenue effects of increasing the nonexclusive good, and (3) the distorting taxes used to raise revenue are optimally set.[6,7]

Concluding Comment

The public good example emphasizes an important yet discouraging point: Even small departures from a first-best environment can generate staggering problems for public sector decision-making. All we did was introduce distorting taxation into an otherwise first-best policy environment. When one considers that the number of real-world distortions is far greater than simply the need to use distorting taxes, the prospects for achieving a unified normative theory of the public sector are indeed discouraging.

THE COASE THEOREM, BARGAINING, AND PRIVATE INFORMATION

The Coase theorem holds out hope that government intervention might not always be required to solve market failures such as externalities. Its premise is that in the absence of private information and transaction costs, private agents have an incentive to bargain with one another and write whatever contracts are necessary to extract all pareto-superior gains and thereby reach the pareto optimum. The only prior requirement is the assignment of property rights, so that ownership of the activities giving rise to or receiving the externalities is clearly established.

No one expected the Coase theorem to apply to instances of widespread externalities because the transaction costs of bringing a large number of agents into a bargaining

4. It may not be, given that factors are subsidized.

5. Atkinson and Stern (1974), p. 123. The analysis of α/λ closely follows their derivation. See Wildasin (1984), for a complete analysis of the effect of the two tax terms on the allocation of the public good. He considers the cases in which some private goods are complements or substitutes to the public good.

6. Karen Conway examined the effect of government expenditures on the supply of labor using a sample of males and females from the 1980 PSID. She achieved variation in government expenditures across individuals by using combinations of state; state and local; and state, local, and federal spending as the government variable. She tried both aggregate expenditures and various individual expenditure categories. The estimates proved to be highly sensitive to sample size and attempts to control for state-fixed effects. Conway concludes that labor supply and government spending are neither ordinary nor compensated independents for men and unmarried women. She could not reject the hypothesis that labor supply and government spending are compensated and ordinary independents for married women. See Conway (1997).

7. Cost–benefit practitioners might consult Sandmo (1998). In this article, Agnar Sandmo presents a simple model of heterogeneous consumers consisting of a composite consumer good, labor, and one nonexclusive public good. Social welfare is utilitarian. He uses the model to demonstrate the effect of four factors on the marginal cost of public funds in line with Eqn (22.11) or (22.15): (1) the sources of tax revenues, particularly whether there is the possibility of a lump-sum tax equal for everyone; (2) whether taxes are set optimally; (3) the distributional characteristic of the public good, defined in terms of the covariance between individuals' MRS and their marginal utility of income; and (4) the effect of changes in the public good on tax revenues. Sandmo argues that (1) and (2) should be considered in calculating a marginal cost of public funds but not (3) and (4), since they are specific to particular public goods.

setting were likely to be formidable. But the hope was that externalities involving only a few agents could be settled efficiently through bargaining rather than by government intervention.

Coase published his theorem in 1960. Since that time, developments in bargaining theory have pretty much dashed the hopes of the theorem even in the case of small numbers of agents. There are two main problems. One is that the formation of coalitions through cooperative bargaining requires that certain conditions be satisfied for the coalitions to be stable. Unfortunately, the set of stable coalitions may not include the pareto-optimal allocation; and even if it does, the bargaining process may not settle on the pareto-optimal coalition. The other problem is that private information may limit the payments that agents are willing to make to other parties as part of a bargain. The acceptable range of payments may preclude the payment that is necessary to achieve the pareto optimum.

The Coase theorem can get around these problems in principle by assuming them away. Rational agents may be seen as rejecting any pareto-inferior coalitions simply because they are pareto inferior. Or, rational agents may be presumed to reveal their private information willingly if doing so would lead to pareto-superior allocations. Assumptions such as these effectively turn the theorem into a tautology. But, if agents act independently and self-interestedly and enter into bargains voluntarily, then the difficulties posed by the requirements of bargaining stability and the presence of private information should not be assumed away. They should be considered in deciding whether government intervention can improve upon a private sector that includes bargaining as well as market exchange.

Bargaining Set Stability and the Coase Theorem

A simple example developed by Varouj Aivasion, Jeffrey Callen, and Irwin Lipnowski illustrates the problems in achieving efficient solutions through cooperative bargaining even in the case of perfect information and small numbers (Aivasian et al., 1987). Suppose there are three agents: two factories (agents 1 and 2) and a laundry (agent 3). Smoke from the factories is an external diseconomy to the laundry. The net values of the factories and laundry, if they act alone or form two-party coalitions or form a three-party coalition, are as follows

Acting Alone	Two-Party Coalitions	Three-Party Coalition
$V(1) = 1$	$V(1, 2) = 8$	$V(1, 2, 3) = 12$
$V(2) = 2$	$V(1, 3) = 9$	
$V(3) = 3$	$V(2, 3) = 10$	

All two-party coalitions improve upon the stand-alone outcomes, always providing a net value of 11 for the three agents combined. The three-party coalition yields the most net value; it is the pareto-optimal solution in this example. A story behind these net values might be that the two factories enjoy synergies if merged and that are absent if they act alone; the merger of either one of the factories and the laundry internalizes the externality, which increases the attainable net value relative to acting alone; and the three-party coalition has the advantage of realizing the factory synergies and internalizing both externalities.

Bargaining Set Stability

Two commonly accepted requirements for stable bargaining sets are individual rationality and coalition stability. *Individual rationality* says that agents will not accept an outcome as part of a coalition that is worse than the outcome they can achieve by acting on their own. Thus, the net values in the first column above set a floor on the values the agents will accept as part of any coalition. *Coalition stability* says that any credible objection to a coalition by one of the members must be able to be met by a credible counterobjection by another member of the collation to ensure that the coalition is stable. A credible objection is an announcement by one member (say, agent *i*) that he or she can form another coalition consisting of himself or herself, some members of the current coalition, and perhaps some agents currently outside the coalition such that he or she is better off under the new coalition and no member of the new coalition is worse off. If this is true, then he will break away and form the new coalition unless someone else can mount a credible counterobjection of the same kind. For example, agent *j* might counterobject that if agent *i* were to do this, he or she could form yet another coalition consisting of all the members in agent *i*'s new coalition, except person *i*, and perhaps some other people such that he or she is better off and no one else is worse off relative to their position with agent *i*'s new coalition. Faced with counterobjections of this kind, no one can gain by breaking away from the coalition and the coalition is stable.

The coalitions that meet the objection/counterobjection test for coalition stability in Aivasion et al.'s example are[8]

	{($Net Values): [Coalition]}
One-agent	{(1; 2; 3): [1, 2, 3]}
Two-agent	{(3.5; 4.5; 3): [(1, 2), 3]}
	{(3.5; 2; 5.5): [(1, 3), 2]}
	{(1; 4.5; 5.5): [(2, 3), 1]}
Three-agent	{(3; 4; 5): [(1, 2, 3)]}

8. The values are the solution to a set of linear inequalities that satisfy the conditions for coalition stability. The equations are in footnote 6 of the A-C-L paper. We will illustrate coalition stability by some examples here.

Consider the first two-agent coalition as an example. Suppose agent 1 objects and proposes to form a new two-agent coalition with agent 3, with the values:

$$\{(3.5 + e; 2; 5.5 - e)\text{: } [(1, 3), 2]\}$$

Agent 2 can counterobject with the following proposed coalition:

$$\{(1; 4.5 + e; 5.5 - e)\text{: } [(2, 3), 1)]\}$$

Thus, each can credibly block the other's attempt to break away from the coalition.

Similarly, any attempt by one of the agents in the three-agent coalition to break away and form a two-agent coalition is subject to a credible counterobjection. For example, if agent 1 objects and wants to break away with agent 3, offering {(3.5 + *e*; 2; 5.5 − *e*): [(1, 3), 2)]}, then agent 2 can counterobject and break away with agent 3, offering {(1; 4.5 + *e*; 5.5 − *e*): [(2, 3), 1]}.

That a three-agent coalition with one division of the combined net value is stable may seem encouraging, but the bargaining process may never get there. One problem is that any pareto-superior move from one of the two-agent coalitions does not produce a stable coalition. Consider the move from the two-agent coalition {(3.5; 4.5; 3): [(1, 2), 3)]} to the pareto-superior three-agent coalition {(3.75, 4.75, 3.5): [(1, 2, 3)]}. Suppose agent 3 objects to the new coalition and wants to join agent 1 in a two-agent coalition, {(3.75 + *e*, 2, 5.25 − *e*): [(1,3), 2)]}. Agent 2 cannot credibly counterobject because with only $8 to split up between agents 1 and 2, it would have to accept {(3.75 + *e*, 4.25 − *e*, 3.5): [(1, 2), 3)]}. The pareto-superior coalition is not stable. Agent 2 might make this counterproposal out of spite, but adding a spite motive makes it unclear what the final bargaining equilibrium might be or, indeed, if any coalition is stable.

An additional problem is that each agent is better off as part of one of the stable two-agent coalitions than in the stable three-agent coalition. Thus, any two agents have a strong incentive to form one of the stable two-agent coalitions as a preemptive move rather than join the three-agent coalition.[9] Therefore, despite the presence of a pareto-optimal and stable three-agent coalition, the bargaining process in this example is highly likely to form a two-agent coalition with a combined net value of 11 rather than 12 for the three-agent coalition.

Economists have proposed a number of different equilibrium concepts in cooperative bargaining settings. There is not one accepted definition of equilibrium. Nonetheless, individual rationality and coalition stability are fairly compelling concepts, and the example of Aivasion et al. shows that imposing them on the bargaining process can undermine the Coase theorem.

9. More generally, the three-agent coalition violates coalition rationality, another widely accepted concept for bargaining set stability. *Coalition rationality* says that any subgroup of agents in a coalition will never accept another coalition in which the total payments to the subgroup are smaller.

Private Information

The existence of private information makes the chances of achieving an efficient bargaining outcome highly unlikely. Even bargains between two agents can fail to achieve an efficient outcome. A two-agent example provided by Peter Klibanoff and Jonathan Morduch illustrates the nature of the problem (Klibanoff and Morduch, 1995).

Suppose that production of firm 1 generates an external economy of size w for firm 2. The externality w is a constant independent of the size of firm 1's output. Firm 2 cannot be certain whether firm 1 will produce. All it knows is that the net value of firm 1 is a random variable, v, ranging from a low value of α to a high value of β, $\alpha < 0$ and $\beta > 0$, with density function $f(v)$ and cumulative density function $F(v)$. Firm 2 is free to engage in a voluntary negotiation with firm 1 and offer a subsidy to encourage firm 1 to produce. The question is how high the subsidy should be.

The pareto-optimal solution is for firm 1 to produce as long as $w + v > 0$. If firm 2 had perfect information about firm 1 and knew that $v < 0$, then it would negotiate a subsidy to firm 1 as long as $w > |v|$. If it knew that $v > 0$, then it would let firm 1 produce without negotiating a subsidy. The pareto-optimal solution obtains in either case.

With uncertainty about v, however, firm 2 has to weigh the expected benefits and costs of any subsidy it gives to firm 1. Suppose it offers a subsidy of x. Then its expected cost is $x(1 - F(-x))$, the subsidy times the probability that firm 1 will produce, given the subsidy. The expected benefit is $w[(1 - F(-x)) - (1-F(0))]$, the value of the externality times the *increase* in the probability that firm 1 will produce, given the subsidy. The increase in the probability is the probability that firm 1 will produce given the subsidy x minus the probability that firm 1 will produce without a subsidy. Thus, firm 2 offers a subsidy if:

$$\begin{aligned} &w[(1 - F(-x)) - (1 - F(0))] \\ &- x[1 - F(-x)] \geq 0, \quad \text{or} \end{aligned} \tag{22.25}$$

$$w[F(0) - F(-x)] \geq x[1 - F(-x)] \tag{22.26}$$

The probability that firm 1 will produce without a subsidy greatly reduces the probability that firm 2 will subsidize firm 1. To see this, assume $f(v)$ is the uniform distribution over the interval $[\alpha, \beta]$, such that

$F(x) = (x - \alpha)/(\beta - \alpha)$. Substituting for $F(-x)$ into Eqn (22.26) yields

$$w[(-\alpha + x + \alpha)/(\beta - \alpha)] \geq x[1 + (x + \alpha)/(\beta - \alpha)] \tag{22.27}$$

Multiplying by $(\beta - \alpha)$ and rearranging terms,

$$wx \geq x(\beta + x) \tag{22.28}$$

$$w \geq \beta + x \tag{22.29}$$

The externality has to be very large, larger than the highest net value of firm 1's production, for firm 2 to subsidize firm 1. Smaller externalities will preclude a subsidy even though $w + v$ might be greater than 0. For example, suppose that $\alpha = -1/2$, $\beta = 1$, and $w = 1$. Since $w = \beta$, there will be no negotiation and subsidy even though a subsidy of $1 > x > 1/2$ would guarantee that firm 1 produces and would make both firms better off. Firm 1 would have positive net value inclusive of the subsidy and firm 2 would enjoy a positive externality net of the subsidy. Private information undermines the Coase theorem.

The government can guarantee the first-best outcome if it makes a side deal with firm 1 that it will not produce unless firm 2 gives it a subsidy of w. This is in effect the Pigovian tax solution, since w is the marginal as well as the total external benefit to firm 2 given that firm 1 is making a produce/do not produce decision. Firm 1 will produce under this subsidy as long as $w > |v|$. Also, firm 2 will agree to pay the subsidy because it knows that firm 1 will not produce without the subsidy. Consequently, the term $(1 - F(0))$ drops out from the expected benefits in Eqn (22.25), and firm 2 offers a subsidy x as long as $w \geq x$. The point is that the Coase theorem can be rescued in the face of private information, but only if the government asserts itself in some coercive fashion. Another example of achieving the first best through government coercion is the Clarke tax scheme, described in Chapter 6, in which the government forces people to participate in order to get them to reveal their demand curves for a nonexclusive good.

If, however, the negotiations remain voluntary and both agents honor the individual rationality condition on bargains, then the first-best outcome is not guaranteed under private information. In the Klibanoff–Morduch example, under the uniform distribution, the first best can be guaranteed only if the externality is so large that firm 2 is willing to ensure that firm 1 produces by giving it a subsidy $x = |\alpha|$, the smallest possible value of v. Plugging $-x = \alpha$ into Eqn (22.26), noting that $w \geq x = -\alpha$ for firm 2 to offer the subsidy, and solving for w:

$$w[F(0) - F(\alpha)] \geq -\alpha[1 - F(\alpha)] \tag{22.30}$$

or

$$w \geq -\alpha/F(0) \tag{22.31}$$

If $-\alpha = \beta$, so that the uniform distribution of v is symmetric around 0, then $F(0) = 1/2$ and

$$w \geq -2\alpha = 2\beta \tag{22.32}$$

The externality has to be greater than *twice* the highest net value of firm 1's production to ensure the first-best solution under private information. This is undoubtedly an unrealistically large externality in most practical applications.

A final question is whether a negotiated subsidy is better than autonomy, even if it is not first best. The answer turns on whether Eqn (22.25) is satisfied for some $v^* < 0$, rewritten here as

$$K(v^*) = w[(1 - F(v^*)) - (1 - F(0)] + v^*[1 - F(v^*)] \geq 0 \tag{22.33}$$

Note that $K(0) = 0$, and

$$dK(v^*)/dv^* = 1 - F(v^*) - (w + v^*)f(v^*) \tag{22.34}$$

The sign of $dK(v^*)/dv^* = \text{sign}\left[\frac{(1-F(v^*))}{f(v^*)} - (w + v^*)\right]$

$\frac{(1-F(v^*))}{f(v^*)}$ is nonincreasing for many distributions, including the uniform distribution. Therefore, a $v^* < 0$ satisfying Eqn (22.25) exists if and only if $dK(v^*)/dv^*|_{v^*=0} \leq 0$, which is equivalent to

$$w \geq \frac{(1 - F(0))}{f(0)} \tag{22.35}$$

For the uniform distribution symmetric around 0,

$$w \geq \frac{(1 - 1/2)}{\frac{1}{2\beta}} = \beta = -\alpha \tag{22.36}$$

We have already seen that firm 2 will offer firm 1 a subsidy as long as $w \geq \beta$, and that the subsidy will be first best if $w \geq 2\beta$. Equation (22.36) establishes that a negotiated subsidy is better than autonomy when $\beta \leq w \leq 2\beta$. It reduces inefficiency, but it is a second-best solution. Finally, the inefficiency of autonomy rises in the range of $0 \leq w \leq \beta$. Figure 22.1 summarizes the outcomes for different values of w (again assuming a uniform distribution of v symmetric around 0).

Market Power and Private Information

Eric Maskin has noted that the bargaining inefficiencies brought on by private information in two-agent externality models, such as the Klibanoff–Morduch model, are inherent in *any* exchange between two or more agents (Maskin, 1994). The externality is not the source of the problem. Instead, the inefficiency arises only because one or more of the parties has market power: In the Klibanoff–Morduch model, firm 2 has the ability to set the subsidy x. If all firms were price takers, then the

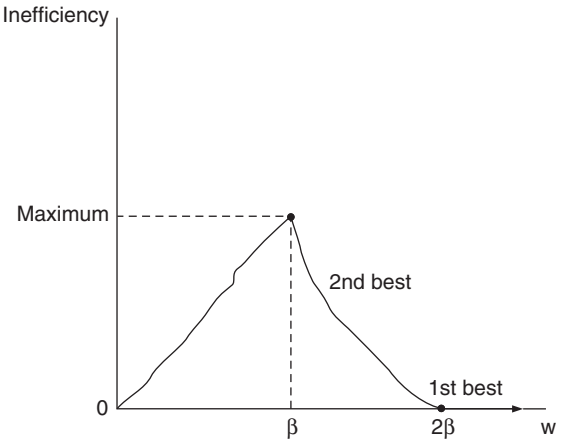

FIGURE 22.1

combination of externalities and private information would not necessarily prevent pareto-optimal bargaining even among a large number of agents.

The essence of the problem when an externality is involved is the following. Suppose the beneficiary of an external economy receives a marginal benefit of w, as in the Klibanoff–Morduch model. If the beneficiary pays the generator of the externality a fee equal to w, the Pigovian tax or pseudo-competitive price, and w exceeds the marginal cost of providing the externality, then the externality will be provided in an efficient manner. If, however, the beneficiary has market power, then it will set its offer below w to try to reap some gain on the margin. The generator of the externality cannot be certain that w is the marginal benefit, so it cannot insist on w in payment. The problem is that the beneficiary's offer may be less than the marginal cost to the generator even if w is not, in which case the externality is not provided and the outcome is not pareto optimal.

Maskin uses a famous externality example of long standing—the beekeepers and the apple growers, first described by James Meade—to illustrate the efficiency of price-taking behavior in the presence of private information. The beekeepers keep bees to produce honey, but the bees provide an external economy to the apple growers by pollinating the apple trees. The problem is to efficiently account for the pollination externality by increasing the number of bees kept by the beekeepers.

To provide a competitive setting, Maskin assumes that there are a large and equal number of beekeepers and apple orchards, n. In the baseline example, each beekeeper is adjacent to one apple orchard, and its pollination externality is experienced only by the adjacent apple orchard. Also, the beekeepers have the power of exclusion. They can prevent the bees from pollinating the apple trees if the price they are offered for the externality is below their marginal cost. The apple orchards are identical, as are the beekeepers.

The beekeepers' costs, c_j, are private information to the beekeepers. c_j is a random variable drawn from the uniform distribution over the interval from 1 to 3. The apple growers' external economy from the number of bees, x, is given by the utility function $U = \theta_i x - x^2$, where θ_i is the private information to the apple growers. θ_i is a random variable also drawn from the uniform distribution over the interval from 1 to 3.

Write the supply and demand functions for bees as $s(c_j, p)$ and $d(\theta_i, p)$. Given a competitively determined price, p, the beekeepers' supply decision for the marginal bee is

$$\begin{aligned} s(c_j, p) &= 1, \quad c_j \le p \\ &= 0, \quad c_j > p \end{aligned} \tag{22.37}$$

The producer surplus of the apple growers from the bees, given the price p, is $U(x) = \theta_i x - x^2 - px$, so that the surplus on the margin is $U' = \theta_i - 2x - p$. $U' = 0$ when $x = (\theta_i - p)/2$. Therefore, their demand for bees is

$$d(\theta_i, p) = \max((\theta_i - p)/2, 0) \tag{22.38}$$

The price is determined in the competitive marketplace. It is the solution to:

$$\frac{1}{n}\sum_{i=1}^{n} d(\theta_i, p^*) = \frac{1}{n}\sum_{j=1}^{n} s(c_j, p^*) \tag{22.39}$$

written from the perspective of each individual apple orchard and beekeeper. For large n, Eqn (22.39) is approximately equivalent to:

$$E[d(\theta_i\, p^*)] = \Pr(s(c_j, p^*) = 1) \tag{22.40}$$

The uniform density function over which θ_i and c_j are drawn is $f(z) = 1/2$. Therefore,

$$\begin{aligned} E[d(\theta_i, p^*)] &= \int_p^3 \left(\frac{\theta_i - p}{2}\right)\cdot\frac{1}{2}d\theta = \Big|_p^3 \frac{(\theta_i - p)^2}{8} \\ &= \frac{9 - 6p - p^2}{8} \end{aligned} \tag{22.41}$$

and

$$\Pr(s(c_j, p^*) = 1) = \int_1^p \frac{1}{2}dc = \Big|_1^p \frac{1}{2}c = \frac{p-1}{2} \tag{22.42}$$

Therefore, p^* is the solution to:

$$\frac{9 - 6p - p^2}{8} = \frac{p-1}{2} \tag{22.43}$$

$$p^* = 5 - 2\sqrt{3} \tag{22.44}$$

Competitive pricing is ex ante efficient assuming that the agents want to maximize their producer surpluses. It calls forth the correct amount of bees even in the presence of

private information.[10] The Coase theorem is vindicated, all the more surprising because the number of agents is large.

This result is tempered by a number of considerations, however. First, the one-on-one nature of the externality makes these transactions not really different in kind from any transaction of a private good between two agents. Second, the ability of the beekeeper to exclude the pollination services of the bees prevents the government from having to assign property rights to the beekeepers. Third, the model ignores transaction costs, which in most many-agent settings are likely to be nontrivial. Finally, the result does not hold if the externalities generated by any one beekeeper extend beyond a single apple orchard.

Nonexclusive Externalities

Maskin develops the final point by considering the nonexclusive case in which the bees can fly to any orchard. Therefore, one bee can be expected to provide $1/n$ units of service to each apple orchard. The efficient solution in this case is that beekeeper j should keep an additional bee as long as

$$\frac{1}{n}\sum_{i=1}^{n}(\theta_i - 2x) \geq c_j \tag{22.45}$$

That is, the sum of the marginal benefits equals the marginal cost.

The LHS of Eqn (22.45) is approximately equal to its expected value for large n. Hence,

$$E(\theta_i) - 2E(x) = 2 - 2(p-1)/2 = 3 - p \tag{22.46}$$

where $E(x)$ is given by (22.42), the expected supply of bees. Hence,

$$3 - p \geq c_j = p \tag{22.47}$$

$$p^* = 3/2 \tag{22.48}$$

Therefore, beekeepers with $c_j < 3/2$ produce a bee, about 1/4 of the beekeepers given that the uniform distribution for c_j is defined over the interval from 1 to 3. The average cost to the beekeepers who keep a bee is the expected value of c_j over the interval of 1 to 3/2, which equals 5/4. Thus, for large n, every apple grower should pay a price of 5/16 $\left(=\frac{1}{4}\cdot\frac{5}{4}\right)$ to ensure the efficient number of bees. But, the free-rider problem prevents this solution: Each apple grower has an incentive not to pay and free ride on the goodwill of the other growers.

The government can enforce the efficient outcome by having agents sign a contract in which: (1) each beekeeper agrees to announce whether it is willing to supply a bee if paid a fee of 3/2; (2) a random selection of (approximately) 3/4 of the apple growers agree to pay a fee of $\frac{1}{2}\left(\frac{1}{2}\cdot\frac{3}{4} = \frac{3}{8} = \frac{3}{2}\cdot\frac{1}{4}\right)$; and (3) nonparticipating apple growers have to pay the fee. This last condition ensures that apple growers will sign the contract: If they sign, they may not be selected to pay the fee, yet they enjoy the same amount of services. One problem with this solution is that the government may not know what the proper fees should be. In any event, this is yet another example of government coercion being required to overcome the free-rider problem with nonexclusive externalities.

Concluding Comments

Although Maskin's analysis clarifies the nature of the bargaining problems in the presence of externalities and private information, it hardly rescues the Coase theorem. On the one hand, when the number of agents is small, some or all of them are likely to have market power. They may not agree to pseudo-competitive fees, in which case the bargains are unlikely to be pareto efficient. On the other hand, externalities that affect a large number of agents are likely to give rise to the free-rider problem. (Transaction costs are also likely to prevent efficient bargains.) The models presented in this section indicate that the solution in either case involves some form of government coercion. The combination of voluntary bargaining, individual rationality, and private information is unlikely to produce pareto-efficient outcomes, contrary to Coase's expectations for bargained solutions to market failures.[11,12]

10. The solution is only approximately efficient since Eqn (22.40) is only an approximation of Eqn (22.39) for large n. Some beekeepers would have to be chosen at random and required either to keep or not to keep an additional bee to satisfy that supply = demand equilibrium.

11. An excellent discussion of this last point can be found in Farrell, 1987. Be mindful, however, that Maskin was responding to Farrell's article in clarifying the source of the bargaining problem with private information.

12. Avinash Dixit and Mancur Olson constructed a simple model to emphasize the difficulties that the combination of voluntary participation and transactions costs causes for the Coase theorem. A group of citizens agree to meet and provide a nonexclusive good if the sum of the benefits of the good to the people at the meeting exceeds its cost. There are no problems reaching agreement at the meeting since there is no private information and no transactions costs to hinder the negotiation of a pareto improvement. The problem is getting people to come to the meeting in the first place, since nonparticipants share the benefits of the good but do not have to pay any of its costs. Asking nonparticipants to pay would be coercive. If the number of citizens exceeds the number required at a meeting to provide the good, then people have a strong incentive to free ride and not attend, and the good may not be provided. Dixit and Olson describe some options in which everyone must participate for the good to be provided that would be pareto optimal, but show that these options are undermined by small transaction costs of attending the meeting (or impatience, if the call to meet can be infinitely repeated should it fail to attract everyone). The Coase theorem rests on shaky grounds even without private information, at least for the provision of nonexclusive goods (Dixit and Olson, 2000).

These difficulties with the Coase theorem notwithstanding, one should not end a discussion of the theorem without giving Coase his due. Coase's insight that the assignment of property rights in situations involving externalities can potentially lead to efficient outcomes is an important one. Among other things, it is transforming the commercial fishing industry.

Fishing stocks were on the verge of collapsing in the last quarter of the twentieth century for a number of commercial fish such as cod and halibut. Governments responded by establishing commissions to regulate the catch in the fishing beds under their jurisdiction in an effort to maintain the stock of each fish at a sustainable level. The fishing for Pacific halibut off the coast of British Columbia is a case in point, although a large number of examples throughout the world would serve just as well.

In 1923, the U.S. and Canadian governments established the International Pacific Halibut Commission (IPHC) to oversee the fishing of halibut. When the stock of halibut off British Columbia became dangerously low in the 1970s, the IPHC established a series of regulations in an effort to keep the stock of halibut sustainable. They began by issuing a fixed number of licenses, 435, one per vessel, to limit the number of vessels that could fish the halibut beds. But technical improvements in long-line gear, which is the primary method used to catch halibut, made the vessels so productive that the stock of halibut continued to decline. The IPHC responded with a limit on the total allowable catch (TAC) by all the vessels combined, a limit designed to keep the remaining stock sustainable. The TAC did not work well at all. Once the fishing season began, each fisher had an incentive to go out every day and bring in as much fish as possible until the TAC was reached. Given the productivity of the vessels, the TAC was reached in 6 days. This meant that the vessels were often going out in rough seas and fishing close together, conditions that made a hazardous occupation much more dangerous. Also, with all the halibut caught within a week, the fishers were forced to sell their fish to the fish processors, which gave the processors monopsony power over the fishers.

Exasperated by the results of the aggregate quotas, the IPHC turned in 1991 to a system of individual transferable quotas (ITQ) that had first been adopted by New Zealand in 1986. Each licensed vessel was given an ITQ that determined the amount of halibut the vessel could bring in during the fishing season. The quota under an ITQ was determined by a formula based on the length of the vessel and its catch in the preceding 4 years. The sum of the individual quotas under the ITQs equaled the sustainable TAC for the halibut bed. In addition, the ITQs were split into two shares and each share could be sold to other vessels, including those already licensed, although no one vessel could hold more than four ITQ shares.

The ITQ approach is very much a Coasian solution, since it essentially gives each fisher a property right over part of the halibut catch. And it had the predictable beneficial results. Fishers no longer had to rush to catch their fish. As a result, the fishing season spread out from six days to many months. Fishing became much safer. The fishers could avoid bad weather and the number of vessels with ITQs decreased 29% from 435 to 309 from 1991 to 1994 as some fishers sold their ITQs to the more efficient vessels and left the business. Also, fishers could bring fresh fish to market over a much longer period of time, which meant they could receive higher prices for fresh fish rather than being forced to sell almost all their catch at much lower prices to the fish processors. The total revenues for the fishers increased by $23 million from 1991 to 1994. The profits would have been even greater had there been no restrictions on the number of ITQ shares one vessel could have, but the IPHC instituted the limit to allay fears that a handful of the largest, most efficient vessels would buy up all the shares and drive the smaller vessels out of business. The goal was to preserve a long-standing way of life for the small fisher. This goal is misguided from a purely economic perspective, however, since if a larger vessel really is more efficient, it can offer the smaller fisher a price for his two ITQ shares that would exceed the present value of the smaller fisher's annual net income from fishing (Grafton et al., 2000, The data on the change in number of vessels and the increase in revenues from 1991 to 1994 is on p. 689.).

In summary, the establishment of marketable property rights to fishing beds is an instance in which assigning property rights to control for an externality (i.e., overfishing) is implementable, and it works about as well as the Coase theorem suggests that it would.

REFERENCES

Aivasian, V., Callen, J., Lipnowski, I., November 1987. The Coase theorem and coalitional stability. Economica 54 (216), 517–520.

Atkinson, A., Stern, N., January 1974. Pigou, taxation, and public goods. Review of Economic Studies 41 (1), 119–128.

Conway, K., February 1997. Labor supply, taxes, and government spending: a microeconometric analysis. Review of Economics and Statistics 79 (1), 50–67.

Diamond, P., November 1975. A many person Ramsey tax rule. Journal of Public Economics 4 (4), 335–342.

Dixit, A., Olson, M., June 2000. Does voluntary participation undermine the Coase theorem? Journal of Public Economics 76 (3), 309–335.

Farrell, J., Fall 1987. Information and the Coase theorem. Journal of Economic Perspectives 1 (2), 113–129.

Grafton, R., Squires, D., Fox, K., October 2000. Private property and economic efficiency: a study of a common-pool resource. Journal of Law and Economics. The data on the change in number of vessels and the increase in revenues from 1991 to 1994 is on p. 689, 43 (2), 679–726.

Klibanoff, P., Morduch, J., April 1995. Decentralization, externalities, and efficiency. Review of Economic Studies 62 (2), 223–248.

Maskin, E., May 1994. The invisible hand and externalities. American Economic Association Papers and Proceedings 84 (2), 333–337.

Pigou, A.C., 1947. A Study in Public Finance, third ed. Macmillan, London.

Sandmo, A., December 1998. Redistribution and the marginal cost of public funds. Journal of Public Economics 70 (3), 365–382.

Wildasin, D., April 1984. On public good provision with distortionary taxation. Economic Inquiry 22 (2), 227–243.

Chapter 23

Decreasing Costs and the Theory of the Second-Best—The Boiteux Problem

Chapter Outline

Chapter 9 developed the three first-best decision rules for decreasing-cost services:

1. *A decreasing-cost industry is a natural monopoly.* It should consist of a single firm to minimize the total cost of producing any given output.
2. *Price must equal marginal cost for pareto optimality.* Achieving this result requires either government regulation or government provision of the service, since a profit-maximizing monopolist would presumably set marginal revenue equal to marginal cost.
3. *Marginal-cost pricing implies operating losses with decreasing unit costs.* Therefore, the government must subsidize the firm's losses with a lump-sum transfer so that the investors can earn a return equal to the opportunity cost of capital. This transfer simply becomes part of the first-best interpersonal equity conditions for optimal income distribution. That is, in satisfying interpersonal equity, the government must collect enough taxes from one subset of consumers to subsidize all decreasing-cost producers as well as provide transfers to the remaining subset of consumers.[1]

The chapter concluded by pointing out that the United States tends to favor average-cost pricing rather than marginal-cost pricing for the decreasing-cost services. The public apparently views average-cost pricing as being fully consistent with the benefits-received principle of public pricing and therefore more equitable. In contrast, mainstream public sector theory has no use for the benefits-received principle as an equity principle.

Chapter 23 extends the analysis of decreasing-cost firms by considering a common property of these firms that Chapter 9 ignored: They are often multiproduct firms that offer a variety of services to different customers (e.g., the U.S. Postal Service and most public utilities).

THE BOITEUX PROBLEM: THE MULTIPRODUCT DECREASING-COST FIRM

We begin with an analysis of the multiproduct, decreasing-cost firm developed by Marcel Boiteux in the 1950s.[2] Boiteux's analysis is one of the seminal contributions to second-best public expenditure theory. He is as closely associated with the decreasing-cost firm as Paul Samuelson is with the nonexclusive public good.

Boiteux considered the optimal pricing and investment rules for multiproduct decreasing-cost monopolies that are required to raise a given amount of revenue. His model is particularly appropriate for the United States. When faced with multiproduct decreasing-cost firms, the

1. A final point common to all first-best expenditure theory is that the government should allow competitive allocations in all other nondecreasing-cost markets. The simple model of Chapter 9 would have had to add one other good to show this formally, but it was clear from the previous analysis of externalities in Chapters 6–8 that marginal-cost pricing of all other goods is pareto optimal.

2. Boiteux (1971). Jacques Dreze presents a useful interpretation of Boiteux's results in Dreze (1964), pp. 27–34. Our analysis closely follows these two papers. We would also recommend Baumol and Bradford (1970) for an excellent intuitive discussion of the Boiteux problem, including its relationship to the optimal tax literature. The article also presents a brief historical account of the early second-best price and tax literature.

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00023-6

U.S. regulatory agencies simply extend their average-cost pricing philosophy to them. They require that the firm's total revenue equal its total cost across all the products in the aggregate rather than for each product individually. The total cost includes an allowable return to capital. Requiring that total revenue equal total cost (or any other arbitrary amount as in the Boiteux model) renders the analysis second best. The firm's total revenue would be less than its total cost in a first-best environment.

Boiteux analyzed this regulatory problem in the context of a many-person, N goods and factors, general equilibrium model in which all other markets are perfectly competitive and the government has the ability to redistribute endowment income lump sum to satisfy interpersonal equity. This is the same as positing a one-consumer-equivalent economy. It highlights the efficiency aspects of the problem.

The Boiteux problem has general interest for public sector economics far beyond the theory of decreasing costs. It stands as the intellectual precursor to a considerable portion of all second-best tax and expenditure theory developed over the past 50 years. For instance, it turns out to be quite similar to the optimal commodity tax problem of Chapters 13 and 14. Boiteux's model can also be used as a basis for developing production decision rules for any public agency subject to a legislated budget constraint, whether or not the agency supplies decreasing-cost services. Since most governments do restrict agencies in this way, the Boiteux analysis obviously has far-reaching practical significance for government policy.

Analytics of the Boiteux Problem

The essence of the Boiteux problem can be described as follows. Let one production sector ("industry") of the economy be under the direct control (or complete regulation) of the government because it exhibits increasing-returns-to-scale production.[3] Assume that this particular government activity employs many inputs and produces many goods and services according to the implicit government production−possibilities relationship:

$$G(Z_1, \ldots, Z_i, \ldots, Z_N) = G(\vec{\mathbf{Z}}) = 0 \tag{23.1}$$

where $\vec{\mathbf{Z}}$ is an $(N \times 1)$ vector of government inputs and supplies, with element Z_i. (The government need not literally employ all inputs and produce all goods and services in the economy, but it is analytically convenient to use the most general formulation possible. Some (perhaps most) of the Z_i will be identically equal to zero for any given application.) Assume further that government production is twice constrained:

1. The government must buy all inputs and sell all outputs at the vector of producer prices, $\vec{\mathbf{p}} = (p_1, \ldots, p_i, \ldots, p_N)$, faced by the economy's perfectly competitive private sector firms. These prices reflect private sector marginal costs (or values of marginal products). Since there is no taxation in this model, $\vec{\mathbf{p}}$ also serves as the vector of consumer prices.
2. Government purchases and sales must satisfy an overall budget constraint of the general form:

$$\sum_{i=1}^{N} p_i Z_i = B \tag{23.2}$$

where B is set by some legislative body. Setting $B = 0$ is the natural interpretation for the regulated decreasing-cost firms in the United States. Revenue from the sale of all government goods and services at actual market prices must equal the total cost of production. This can be thought of as the average- or full-cost pricing philosophy applied to a multiservice firm (with the allowable return to investors set at the opportunity cost of capital).

The problem, then, is to derive optimal production decision rules for the government control variables, $\vec{\mathbf{Z}}$, given the government's production function and its self-imposed budget constraint.

This problem can be analyzed quite easily using the loss-minimization technique. Assume a one-consumer (equivalent) economy in which all relevant information about the consumer is summarized by the expenditure function:

$$M(\vec{\mathbf{p}}; \overline{U}) = \sum_{i=1}^{N} p_i X_i^{\text{comp}}(\vec{\mathbf{p}}; \overline{U}) \tag{23.3}$$

Assume, further, that private production exhibits general technology with constant returns to scale (CRS), summarized by the profit function:

$$\pi(\vec{\mathbf{p}}) = \sum_{i=1}^{N} p_i y_i(\vec{\mathbf{p}}) \tag{23.4}$$

Since production occurs in the private and public sectors, market clearance must also be specified as[4]

$$M_i(\vec{\mathbf{p}}; \overline{U}) = \pi_i(\vec{\mathbf{p}}) + Z_i \quad i = 2, \ldots, N \tag{23.5}$$

3. As will become evident, the increasing-returns-to-scale assumption merely provides a convenient motivation for government regulation or control. It is not a necessary condition for any of the theorems derived in this chapter.

4. Recall that $M_i = X_i^{\text{comp}}(\vec{\mathbf{p}}; \overline{U})$ and $\pi_i = Y_i(\vec{\mathbf{p}})$.

Recall that all markets cannot clear in terms of compensated demand (supply) functions. Therefore, let the first market remain uncleared, with compensation occurring in terms of good 1. The first good also serves as the numeraire ($p_1 \equiv 1$). Finally, as a matter of convenience, define the government's production function such that $\partial G/\partial Z_1 = G_1 = 1$, or $Z_1 = -g\ (Z_2, \ldots, Z_N)$, with inputs measured negatively. ($\partial g/\partial Z_k \equiv g_{Z_k}$ measures the marginal product of Z_k in producing Z_1 as a positive number.)[5]

Loss is defined as the lump-sum income required to keep the consumer at utility level $\overline{U}$ less all sources of lump-sum income for any given values of the Z_i. In general,[6]

$$L(\vec{\mathbf{Z}}) = M(\vec{\mathbf{p}}; \overline{U}) - \pi(\vec{\mathbf{p}}) - \sum_{i=1}^{N} p_i Z_i - B \tag{23.6}$$

Loss must be minimized with respect to Z_i, subject to the constraints that $\sum_{i=1}^{N} p_i Z_i = B$ and government production. Formally:[7]

$$\begin{aligned} &\min_{(Z_i)} M(\vec{\mathbf{p}}; \overline{U}) - \pi(\vec{\mathbf{p}}) - \sum_{i=1}^{N} p_i Z_i - B \\ &\text{s.t.} \quad \sum_{i=1}^{N} p_i Z_i = B \\ &Z_1 = -g(Z_2, \ldots, Z_N) \end{aligned}$$

Alternatively, directly incorporating the government production function into the government budget constraint,

$$\begin{aligned} &\min_{(Z_i)} M(\vec{\mathbf{p}}; \overline{U}) - \pi(\vec{\mathbf{p}}) + p_1 g(\vec{\mathbf{Z}}) - \sum_{i=2}^{N} p_i Z_i - B \\ &\text{s.t.} \quad -p_1 g(\vec{\mathbf{Z}}) + \sum_{i=2}^{N} p_i Z_i = B \end{aligned}$$

The corresponding Lagrangian equation is

$$\begin{aligned} &\min_{(Z_i)} M(\vec{\mathbf{p}}; \overline{U}) - \pi(\vec{\mathbf{p}}) + p_1 g(\vec{\mathbf{Z}}) - \sum_{i=2}^{N} p_i Z_i - B \\ &\qquad + \lambda \left[-p_1 g(\vec{\mathbf{Z}}) + \sum_{i=2}^{N} p_i Z_i - B \right] \end{aligned}$$

The producer prices, p_i, are functions of Z_i with general technology. Therefore, the first-order conditions with respect to the Z_k are (with $p_1 = 1$):

$$\begin{aligned} &\sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial Z_k} - \sum_{i=2}^{N} \pi_i \frac{\partial p_i}{\partial Z_k} + p_1 g_{Z_k} - \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial Z_k} - p_k \\ &\qquad + \lambda \left[-p_1 g_{Z_k} + \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial Z_k} + p_k \right] = 0 \\ &k = 2, \ldots, N \end{aligned} \tag{23.7}$$

From market clearance,

$$M_i = \pi_i + Z_i \quad i = 2, \ldots, N \tag{23.8}$$

Multiply Eqn (23.8) by $\partial p_i/\partial Z_k$ and sum over all $(N-1)$ equations to obtain:

$$\sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial Z_k} = \sum_{i=2}^{N} \pi_i \frac{\partial p_i}{\partial Z_k} + \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial Z_k} \tag{23.9}$$

Thus, Eqn (23.7) simplifies to

$$\begin{aligned} &p_1 g_{Z_k} - p_k + \lambda \left(-p_1 g_{Z_k} + \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial Z_k} + p_k \right) = 0, \\ &\qquad k = 2, \ldots, N \end{aligned} \tag{23.10}$$

or

$$(\lambda - 1)\left(-p_1 g_{Z_k} + p_k \right) + \lambda \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial Z_k} = 0 \quad k = 2, \ldots, N \tag{23.11}$$

To interpret Eqn (23.11), use the market clearance equations to substitute out the price derivatives, $\partial p_i/\partial Z_k$. Differentiate each of the market clearance equations, Eqn (23.8), with respect to Z_i to obtain

$$\sum_{j=2}^{N} (M_{ij} - \pi_{ij}) \frac{\partial p_j}{\partial Z_k} = a_{ik} = \begin{matrix} \to 0, & i \neq k \\ \to 1, & i = k \end{matrix} \tag{23.12}$$

Differentiating the market clearance relationships with respect to all other Z_i, $i = 2, \ldots, k-1, k+1, \ldots, N$; and writing the results in matrix notation yields

$$(M_{ij} - \pi_{ij}) \left(\frac{\partial p}{\partial Z} \right) = I \tag{23.13}$$

where

M_{ij} = an $(N-1) \times (N-1)$ matrix of derivatives $\partial M_i/\partial p_j$.
π_{ij} = an $(N-1) \times (N-1)$ matrix of derivatives $\partial \pi_i/\partial p_j$.
$\partial p/\partial Z$ = an $(N-1) \times (N-1)$ matrix of derivatives $\partial p_i/\partial Z_j$.
I = the $(N-1) \times (N-1)$ identity matrix.

5. Since Z_1 and Z_k can be either goods or factors, g_{Z_k} can also be interpreted as a technical rate of substitution or a marginal rate of transformation.

6. With CRS production ($\pi(\vec{\mathbf{p}}) = 0$) and the requirement that $\sum_{i=1}^{N} p_i Z_i = B$, the loss function can be simplified to $L(\vec{\mathbf{Z}}) = M(\vec{\mathbf{P}};\ \overline{U})$. The expanded version of loss will be maintained for generality in deriving the optimal production and pricing rules.

7. Following the practice in Chapter 14, the market clearance equations will be kept outside the loss-minimization framework. In this example, they solve for the prices $\vec{\mathbf{p}}$ once the loss-minimizing $\vec{\mathbf{Z}}$ has been determined. Also, recall that $p_1 = 1$.

Thus,

$$\frac{\partial p}{\partial Z} = \left(M_{ij} - \pi_{ij}\right)^{-1} \tag{23.14}$$

Using Eqn (23.14), the entire set of first-order conditions, Eqn (23.11), can be expressed in matrix notation as

$$(\lambda - 1)(-p_1 g_Z + p) + \lambda(M_i - \pi_i)\left(M_{ij} - \pi_{ij}\right)^{-1} = 0 \tag{23.15}$$

where

M_i is the $1 \times (N-1)$ vector $(M_2, \ldots, M_N)$,
π_i is the $1 \times (N-1)$ vector $(\pi_2, \ldots, \pi_N)$,
$(M_i - \pi_i) = Z_i$, and
g_z = the $1 \times (N-1)$ vector $(g_{Z_2}, \ldots, g_{Z_N})$.

Multiplying Eqn (23.15) by $(M_{ij} - \pi_{ij})$ yields:

$$(\lambda - 1)(-p_1 g_Z + p)(M_{ij} - \pi_{ij}) + \lambda(M_i - \pi_i) = 0 \tag{23.16}$$

Select the kth relationship from Eqn (23.16):

$$(\lambda - 1)\sum_{i=2}^{N}(-p_1 g_{Z_i} + p_i)(M_{ik} - \pi_{ik}) + \lambda(M_k - \pi_k) = 0 \tag{23.17}$$

Rearranging terms:

$$\frac{\sum_{i=2}^{N}(-p_1 g_{Z_i} + p_i)(M_{ik} - \pi_{ik})}{M_k - \pi_k} = \frac{-\lambda}{\lambda - 1} \quad k = 2, \ldots, N \tag{23.18}$$

where the right-hand side is a constant independent of k. Written in this form, the first-order conditions can be given an interpretation remarkably similar to the optimal commodity tax rules of Chapters 13 and 14.

As originally defined, the problem asks us to interpret the first-order conditions of Eqn (23.11) as decision rules for the government production variables, Z_i, that is, as government "investment" rules. Using Eqn (23.18), they can also be given a pricing interpretation if one thinks of the government as making "competitive" production decisions in the usual manner. Given a production function $Z_1 = -g(Z_2, \ldots, Z_N) = 0$ and a vector of fixed *shadow* prices for the inputs and outputs $\theta = (\theta_2, \ldots, \theta_i, \ldots, \theta_N)$, a profit-maximizing firm equates $g_i/g_j = \theta_i/\theta_j$, for $i, j = 2, \ldots, N$. Furthermore, if the shadow prices reflect true social opportunity costs for the inputs and outputs, the firm's decision rule is pareto optimal. Condition (23.18) describes, in effect, how to define the optimal shadow prices for the government sector. To see this, let $\vec{\mathbf{p}} = \vec{\boldsymbol{\theta}} + \vec{\mathbf{t}}$, with elements p_i, θ_i, and t_i, respectively, where $\vec{\mathbf{p}}$ is the vector of actual market prices, $\vec{\boldsymbol{\theta}}$ is the vector of optimal shadow prices, and $\vec{\mathbf{t}}$ is a vector of *implicit* taxes driving a wedge between the two sets of prices. Given our normalization, $p_1 \equiv \theta_1$, $t_1 = 0$. Substituting for the p_i in Eqn (23.18),

$$\frac{\sum_{i=2}^{N}(\theta_i + t_i - p_1 g_{Z_i})(M_{ik} - \pi_{ik})}{M_k - \pi_k} = \frac{-\lambda}{\lambda - 1} = C \quad k = 2, \ldots, N \tag{23.19}$$

But, if the government sector is using the θ_i as shadow prices,

$$\theta_i = p_1 g_{Z_i} \quad i = 2, \ldots, N \tag{23.20}$$

Equation (23.20) says, for example, that the government producer hires an input until the value of its marginal product just equals its shadow price. Substituting Eqn (23.20) into (23.19) yields

$$\frac{\sum_{i=2}^{N} t_i(M_{ik} - \pi_{ik})}{M_k - \pi_k} = C \quad k = 2, \ldots, N \tag{23.21}$$

which is virtually identical to the optimal commodity tax rule of Chapters 13 and 14.

Equation (23.21) says that the government should define a new set of shadow prices for use in production decisions by establishing a set of implicit taxes having the following properties (in Boiteux's words): "(The taxes) are proportionate to the infinitesimal variations in price, that, when accompanied by compensating variations in incomes, entail the same proportional change in the demands (supplies) of the goods produced (consumed) by the nationalized sector."[8] Boiteux's interpretation follows directly from the symmetry of the demand and production derivatives, $M_{ik} = M_{ki}$ and $\pi_{ik} = \pi_{ki}$, so that Eqn (23.21) can be rewritten as

$$\frac{\sum_{i=2}^{N} t_i(M_{ki} - \pi_{ki})}{M_k - \pi_k} = C \quad k = 2, \ldots, N \tag{23.22}$$

M_{ki} gives the change in the compensated demand (supply) for good (factor) k in response to a change in the ith consumer price. Similarly, π_{ki} gives the change in supply (demand) of good (factor) k in private production in response to a change in the ith price. Consequently, $(M_{ki} - \pi_{ki})$ gives the change in government supply (demand) of good (factor) k required to maintain compensated market clearance in response to a change in the ith price. The denominator, $(M_k - \pi_k) = Z_k$ from market clearance. Thus, Eqn

8. Boiteux (1971), p. 230. Equation (23.21) also points out that our formulation of $G()$ implies that the government retains control over all prices in the economy, an assumption we have been using all along. If the government is constrained from changing some distorted price–cost margins in the private sector, these additional constraints would change the optimal decision rules, both here and elsewhere in the text. Formally, the new constraints could be represented as $q_i = k_i p_i$, for some i, with k_i constant for good i, and they would have Lagrangian multipliers associated with them in the loss-minimization problem.

(23.22) gives the familiar equal percentage rule, except that the conditions apply only to percentage changes entirely within the government sector.

The crux of the matter, then, can be viewed as defining a correct set of shadow prices on which to base standard "competitive" government production decisions. This suggests that the original problem could have been formulated as a tax-price problem rather than as a quantity problem. Viewed in this way, the problem is indistinguishable from the problem of designing a set of optimal commodity taxes on part of production, to be paid by the producer. The partial tax problem was first described for a single tax in the discussion of corporate tax incidence in Chapter 16 using the two-good, two-factor Harberger model. It can be easily generalized for $(N-1)$ goods and factors by dividing the profit function into two sectors, one taxed and the other untaxed, and using the loss-minimization technique. Consumer and producer prices would differ only in the taxed sector. In the Boiteux problem, the taxes $\vec{\mathbf{t}}$ are implicit and $\vec{\boldsymbol{\theta}}$ are shadow prices. They are not actually observed in the market. In contrast, the taxes in the partial tax problem are real, so that $\vec{\boldsymbol{\theta}}$s define observed gross-of-tax prices (for factors) or net-of-tax prices (for goods) to the firm.

Public Agencies and Private Markets

Whether or not these implicit valuations in the Boiteux formulation affect actual market prices depends upon the relationship of the government producer to the entire market. There are a number of possibilities. Suppose, for example, that the government is merely one of thousands of firms hiring a particular factor of production. It would then be reasonable to assume that its implicit tax had no effect on actual market prices, a situation depicted in Fig. 23.1. The government sets an implicit tax t_k on the purchase of X_k, which drives the factor's shadow price to X_k, from p_k^0 to θ_{gk}^F. However, because it is small relative to the total market of X_k, the price of X_k remains at p_k^0 for all other firms and all consumers.

In fact, Fig. 23.1 is misleading because the government need not design implicit taxes in markets for which the tax does not affect market prices. Consider the kth equation of Eqn (23.11). If Z_k is "small" relative to the entire market for good (factor) k such that $\partial p_i/\partial Z_k = 0$, for $i = 2, \ldots, N$, then Eqn (23.11) becomes

$$(\lambda - 1)\left(-p_1 g_{Z_k} + p_k\right) = 0 \qquad (23.23)$$

which is satisfied if $t_k = 0$. In other words, the government should use the actual market price of p_k in deciding how much Z_k to employ (supply).

Suppose, however, that the government is the only supplier of a particular output X_j. In this case, the implicit tax is virtually identical to a real tax. Refer to Fig. 23.2. The shadow price θ_j equals the firm's actual marginal costs at X_j^F, but because of the implicit tax, the firm charges the consumer p_j^F, equal to measured marginal costs plus the implicit tax t_j. Thus, the consumer is indifferent between the implicit tax or a real partial tax (with lump-sum return

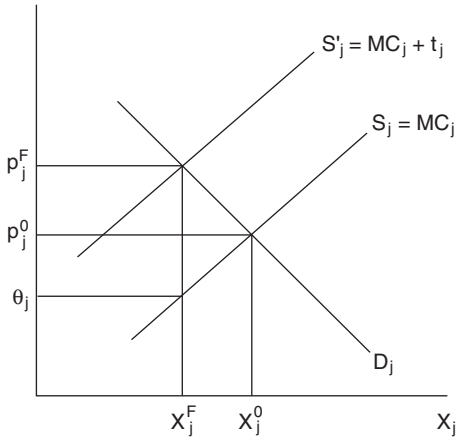

FIGURE 23.2

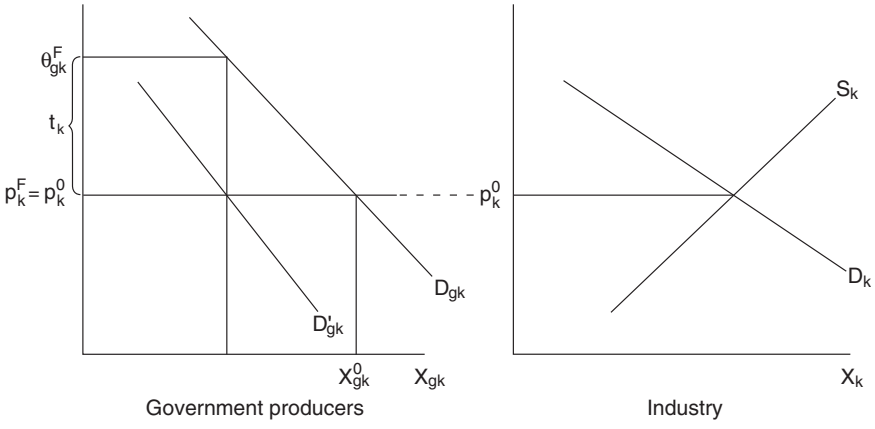

FIGURE 23.1

of the revenues). (Notice that, although the firm receives p_j^F for each unit, it pretends it is receiving only θ_j for the purposes of implicit "competitive" profit maximization.)

These results indicate that the first-order conditions for optimal implicit taxes, Eqn (23.22), may be much easier to approximate than it appears at first pass, certainly much easier than the optimal tax rules for the economy as a whole. In many cases, the government producer will be supplying a few services that are unique to it and buying general factors whose prices are set in large national markets. Thus, it needs only determining implicit taxes on its services. In effect, then, Eqn (23.22) tells the government how to raise prices above measured marginal costs on each of its services in order to satisfy an overall budget constraint. Viewed in this way, Eqn (23.22) provides an efficient second-best algorithm for applying full or "average cost" pricing to the multiproduct firm.

The U.S. Postal Service

The full-cost interpretation of the Boiteux tax rules found its way into the setting of U.S. postal rates, in the form of the inverse elasticity rule (IER). The Postal Service delivers four main classes of mail: first class (letters and post cards); second class (primarily magazines); third class (circulars and other forms of advertising); and fourth class (parcel post). The rates on each class of mail are adjusted periodically in an administrative proceeding presided over by an administrative law judge. The Postal Service proposes a new set of rates to the judge, who then receives testimony from the Postal Service and other interested parties before determining the final rate changes.

The Postal Service claims that its operation exhibits substantial economies of scope: That is, the total cost of delivering all four classes of mail is less than the combined total costs of delivering each class of mail under separate enterprises. The claim is controversial but difficult to prove or disprove because a large component of Postal Service costs is common to more than one class of service. In any event, the Postal Service argues that the costs that can be attributed to each class of mail are much less than the full operating cost of the Postal Service. Therefore, it has to propose rates on each class of mail that will be sufficient to cover its full operating cost, consistent with the structure of the Boiteux problem.

The Postal Service introduced the IER in the 1974 hearings as justification for its rate proposals. In particular, it sought a relatively large increase in the first-class rate and a small increase in the parcel-post rate and argued that its request was efficient because it was in line with the IER. The demand for first-class mail was relatively inelastic. Congress had granted the Postal Service a monopoly on delivering first-class mail and it had no serious competitors at the time. There were no fax machines or e-mails. In contrast, the demand for parcel post was relatively elastic because it faced stiff competition from United Parcel Service (UPS). The IER did not carry the day in 1974, but it was used as the basis for setting rates on the four classes of mail in the 1977 hearings.

UPS always objected strenuously to the use of the IER. It argued that the Postal Service seriously underattributed costs to the various classes of mail, especially to parcel post, and was in effect using the IER to cross-subsidize parcel post with its revenues from first-class mail.[9] The opposition to the IER finally proved persuasive to the administrative law judge. In 1980, the Judge abandoned explicit use of the IER for setting postal rates.

Frank Scott undertook an independent study of the costs of the four classes of mail in 1980 (Scott, 1986). Based on his cost and demand estimates, he determined the rates that were consistent with the IER. He concluded that only the first-class rate was approximately equal to the IER rate, whereas the second-class rate was close to marginal cost, the third-class rate was above the IER rate, and the fourth-class rate was below the IER rate. He estimated that aggregate consumer surplus would have been $14.5 million higher had all the rates been at IER levels, a rather modest increase. The gain in moving to the IER rates was small because the first-class rate was already at the IER rate and first-class mail accounted for 65% of the total revenues of the Postal Service in 1980. Of course, Scott's conclusions are only as accurate as his cost and demand estimates. As noted above, attributing costs to the various classes of mail requires a fair amount of judgment because of the large amount of common costs.

CONSTRAINED GOVERNMENT AGENCIES

Although Boiteux's analysis was motivated by an attempt to develop optimal second-best rules for public monopolies, the resulting decision rules, Eqn (23.11) or (23.22), are directly applicable to any government agency subject to a legislated budget constraint. That this is so is obvious from our original formulation of the Boiteux problem, in which the government is constrained to meet a given revenue target B from the purchase and sales of inputs and outputs at actual market prices $(p_1, \ldots, p_i, \ldots, p_N)$, or $\sum_{i=1}^{N} p_i Z_i = B$. The government has a production function, $Z_1 = -g(Z_2, \ldots, Z_N)$, but there are no formal restrictions on $g(\vec{\mathbf{Z}})$ (other than that it be continuous and twice differentiable). It does not even have to be a homogeneous

9. The IER was first proposed by William Vickrey in the 1974 hearings as a means of determining postal rates on the various classes of mail (Docket No. R74−1). Also see "Postal Rate and Fee Increases," Docket No. R75−1.

function. Clearly, then, the original formulation is a fairly general problem covering any constrained public agency engaged in the production of goods and services. Condition (23.22) suggests that constrained agencies should follow standard competitive production decision rules, based on shadow prices determined by the solution of Eqn (23.22). Once again, the number of shadow prices to be determined depends upon the importance of the agency relative to the national markets for its outputs and inputs. Finally, Boiteux was able to prove as an extension of his results that if there is more than one such constrained sector (agency), each sector has its own set of rules similar to Eqn (23.22) (Boiteux, 1971). This result is obviously of some importance since most government agencies operate under imposed budget constraints, but we will not exhibit it here.

REFERENCES

Baumol, W., Bradford, D., June 1970. Optimal departures from marginal cost pricing. American Economic Review 60 (3), 265–283.

Boiteux, M., September 1971. On the management of public monopolies subject to budgetary constraints. Journal of Economic Theory 3 (3), 219–240 (French in *Econometrica*, January 1956, Trans.).

Dreze, J., June 1964. Some post-war contributions of French economists to theory and public policy. American Economic Review (suppl.) 54 (4-Supplement), 1–64.

Postal Rate and Fee Increases, *Docket No. R75–1; Docket No. R74–1*.

Scott Jr., F., March 1986. Assessing USA postal ratemaking: an application of Ramsey prices. The Journal of Industrial Economics 34 (3), 279–290.

Chapter 24

General Production Rules in a Second-Best Environment

Chapter Outline

This chapter concludes our survey of second-best public expenditure theory by exploring some fairly general propositions about government production in an environment made second best because of distorting taxation. A major goal of the chapter is to integrate our previous results on second-best tax theory with second-best public expenditure theory. Therefore, all the analysis in this chapter employs essentially the same set of assumptions regarding government activity and the underlying structure of the private sector that we have been using time and again.

Regarding the government sector, the government is making a set of production decisions under two constraints: (1) it must buy inputs and sell outputs at the established private sector producer prices and (2) it must cover any resulting deficit (surplus) with distorting commodity taxes levied on the consumer. Otherwise, government production is fully general. The government may buy or sell any inputs or outputs, including those traded in the private sector, and there are no restrictions on the form of the aggregate government production function other than the exclusion of externalities. Following Chapter 21, the government's production function is specified as $G(Z) = 0$, or $Z_1 = -g\ (Z_2, \ldots, Z_N)$, where Z spans (potentially) the entire set of the economy's inputs and outputs. The only difference between the specification of the government sector in this chapter and the specification employed in the Boiteux's analysis is that the government taxes (subsidizes) all consumer transactions to cover its deficits (surpluses), not only those between the consumers and the government.

Regarding the private sector, all markets are assumed to be perfectly competitive and private production exhibits general technology with constant returns to scale (CRS). There can be no pure profits or losses from private production. We also assume that the consumers have no other sources of lump-sum income; all income derives from the sale of variable factors. These assumptions about the private sector are not necessary, but they greatly facilitate the analysis. In sum, the only distortions in the economy that render the analysis second best are the distorting commodity taxes used to cover government production deficits.

Given this analytical framework, the first problem to be considered is the so-called Diamond–Mirrlees problem, which Peter Diamond and James Mirrlees set out in their two-part article in the 1971 *American Economic Review* entitled "Optimal Taxation and Public Production."(Diamond and Mirrlees, 1971) By 1968, when their paper was drafted and widely circulated, the optimal tax rule for a one-consumer (equivalent) economy was well known, but only under the assumption that the government simply raised revenue to be returned lump sum to the consumer. Diamond and Mirrlees added government

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00024-8

production to the standard second-best general equilibrium tax model and asked two questions:

1. How does the existence of government production affect the optimal tax rule? In particular, if the revenue is raised to cover a government production deficit under the conditions set forth above, what form do the tax rules take? They found that the optimal tax rule was unchanged. This result could have been anticipated since it was well known by then that the tax rules as originally derived did not contain any production terms even when *private* production exhibited general technology.
2. Turning the question around: What effect does distorting taxation have on government production rules? Their answer to this question most definitely was unanticipated. They proved that as long as the taxes are set optimally, the government should follow the standard *first-best* production rules, using the private sector producer prices and equating these price ratios to marginal rates of transformation. Distorting taxation necessarily forces society underneath its utility–possibilities frontier, but it should remain on the production–possibilities frontier. This is one of the strongest results in all of second-best theory.

Having established the Diamond and Mirrlees production result, the chapter then generalizes their analysis to consider government production rules under conditions of nonoptimal distorting taxation. As one might suspect, production efficiency no longer holds. In fact, the production rules become fairly complicated. This is especially unfortunate because real-world taxes are likely to be far from optimal.

Any analysis incorporating both second-best tax and expenditure theory is bound to be complex, although the assumptions on the private sector help somewhat in simplifying the analysis. To simplify even further and highlight the efficiency aspects of taxation and government production, we begin the analysis in the context of a one-consumer (equivalent) economy using the technique of loss minimization. We will then conclude the chapter by reworking one of the production exercises in a many-person economy to suggest how equity considerations modify the one-consumer rules.[1]

THE DIAMOND–MIRRLEES PROBLEM: ONE-CONSUMER ECONOMY

Let us establish the general equilibrium framework for the Diamond–Mirrlees problem with some care since we will be using this analytical structure throughout most of the chapter.

The Private Sector

The private sector consists of a single consumer and a set of perfectly competitive producers with general technologies and CRS production. Loss minimization requires that the consumer's decisions be represented by the expenditure function:

$$M\ \left(\vec{\mathbf{q}};\overline{U}\right)\ =\ \sum_{i=1}^{N} q_i X_i^{\text{comp}}\ \left(\vec{\mathbf{q}};\overline{U}\right) \tag{24.1}$$

where

$\vec{\mathbf{q}}$ is the $(N \times 1)$ vector of consumer prices with element q_i (gross of tax for outputs, net of tax for inputs).

$\vec{\mathbf{X}}^{\text{comp}} = \vec{\mathbf{M}}_i$ is the $(N \times 1)$ vector of the compensated demands for goods and factor supplies, with element X_i^{comp} (or M_i).

Let private production be represented by an aggregate profit function:

$$\pi\left(\vec{\mathbf{p}}\right)\ =\ \sum_{i=1}^{N} p_i Y_i\ \left(\vec{\mathbf{p}}\right) \tag{24.2}$$

where

$\vec{\mathbf{p}}$ = the $(N \times 1)$ vector of producer prices with element p_i (gross of tax for inputs, net of tax for outputs).

$\vec{\mathbf{Y}}\ =\ \vec{\pi}_i$ is the $(N \times 1)$ vector of private supplies and factor demands, with element Y_i (or π_i).

with CRS, $\pi(\vec{\mathbf{p}}) \equiv 0$.

The Government Sector

The government has an $(N \times 1)$ vector of production decision variables, $\vec{\mathbf{Z}}$, with element Z_i, related by the aggregate government production function:

$$G\left(\vec{\mathbf{Z}}\right)\ =\ 0\quad \text{or}\quad Z_1\ =\ -g\left(Z_2,\ldots,Z_N\right) \tag{24.3}$$

Since it buys and sells at the private producer prices, the resulting deficit (surplus) from government production is

$$D\ =\ \sum_{i=1}^{N} p_i Z_i \tag{24.4}$$

with inputs measured negatively following the usual convention. The government covers the deficit (surplus) by using an $(N \times 1)$ vector of unit "commodity" taxes $\vec{\mathbf{t}}$, with element t_i, placed on the consumer, such that

1. These analyses draw heavily from two papers by Robin Boadway: Boadway (1975), Boadway (1976). We also benefited from a set of unpublished class notes provided by Peter Diamond.

$\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$. The revenue raised at the compensated equilibrium[2] is $\sum_{i=1}^{N} t_i M_i$, so that the government's budget constraint has the form:

$$\sum_{i=1}^{N} t_i M_i + \sum_{i=1}^{N} p_i Z_i = B \tag{24.5}$$

If B is not equal to zero, the resulting surplus or deficit is returned lump sum to the consumer.

Market Clearance

With two sources of production and general production technology, market clearance must be introduced explicitly into the analysis. We know, however, that all markets cannot clear at the compensated with-tax equilibrium.[3] Therefore, specify

$$M_i\left(\vec{\mathbf{q}};\overline{\mathbf{U}}\right) = \pi_i\left(\vec{\mathbf{p}}\right) + Z_i \quad i = 2,\ldots,N \tag{24.6}$$

and assume that compensation occurs in terms of good 1. Let good 1 also serve as the untaxed numeraire, so that $q_1 \equiv p_1 \equiv 1; t_1 = 0$. This completes all the relevant elements of the general equilibrium framework. The government has $(2N-2)$ control variables at its disposal, $(Z_2, \ldots, Z_N)$ and $(t_2, \ldots, t_N)$.

Loss Minimization

The loss function that the government minimizes with respect to these control variables has the general form:

$$\begin{aligned} L\left(\vec{\mathbf{t}};\vec{Z}\right) &= M\left(\vec{\mathbf{q}};\overline{U}^0\right) - \sum_{i=2}^{N} t_i M_i \\ &\quad - \sum_{i=1}^{N} p_i Z_i - \pi\left(\vec{\mathbf{p}}\right) \end{aligned} \tag{24.7}$$

where, with general technology,

$$\vec{\mathbf{q}} = q\left(\vec{\mathbf{t}};\vec{\mathbf{Z}}\right) \quad \text{and} \quad \vec{\mathbf{p}} = p\left(\vec{\mathbf{t}};\vec{\mathbf{Z}}\right) \tag{24.8}$$

Loss equals the lump-sum income required to keep the consumer indifferent to the gross of tax prices less all sources of lump-sum income resulting from decisions on the government control variables. In this model, lump-sum income derives from two sources: (1) pure economic profits (losses) from private production and (2) the remaining government surplus after taxes have been collected. With CRS, the profit term need not be included. Similarly, if tax revenues just cover government production deficits, the second and third terms could be dropped as well, with loss defined simply as $M(\vec{\mathbf{q}};\overline{U}^0)$. In the interest of generality, however, all these terms are retained in the subsequent analysis.

Loss is minimized subject to government production technology and the government budget constraint. Formally, the problem is

$$\begin{aligned} &\min_{(\vec{\mathbf{t}},\vec{\mathbf{Z}})} M\left(\vec{\mathbf{q}};\overline{U}^0\right) - \sum_{i=2}^{N} t_i M_i - \sum_{i=1}^{N} p_i Z_i - \pi\left(\vec{\mathbf{p}}\right) \\ &s.t. \sum_{i=2}^{N} t_i M_i + \sum_{i=1}^{N} p_i Z_i = B \\ &Z_1 = -g(Z_2,\ldots,Z_N) \end{aligned}$$

Incorporating the government production constraint directly into the analysis and noting that $\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$, the problem can be restated as

$$\begin{aligned} &\min_{(\vec{\mathbf{t}},\vec{\mathbf{Z}})} M\left(\vec{\mathbf{p}} + \vec{\mathbf{t}};\overline{U}^0\right) - \sum_{i=2}^{N} t_i M_i + p_1 g\left(Z_2,\ldots,Z_N\right) \\ &\quad - \sum_{i=2}^{N} p_i Z_i - \pi\left(\vec{\mathbf{p}}\right) \\ &s.t. \sum_{i=2}^{N} t_i M_i - p_1 g\left(Z_2,\ldots,Z_N\right) + \sum_{i=2}^{N} p_i Z_i = B \end{aligned}$$

Also, $q_1 \equiv p_1 \equiv 1$ for $t_1 = 0$, so that there are $(2N-2)$ control variables, $(t_2, \ldots, t_N)$ and $(Z_2, \ldots, Z_N)$.

Finally, the market clearance equations are used to simplify the first-order conditions. Formally, they solve for $\vec{\mathbf{p}}$ given the solution for $\vec{\mathbf{t}}$ and $\vec{\mathbf{Z}}$. The Lagrangian equation for this problem is

$$\begin{aligned} &\min_{(\vec{\mathbf{t}},\vec{\mathbf{Z}})} L = M\left(\vec{\mathbf{p}} + \vec{\mathbf{t}};\overline{U}^0\right) \\ &\quad - \sum_{i=2}^{N} t_i M_i + p_1 g(Z_2,\ldots,Z_N) \\ &\quad - \sum_{i=2}^{N} p_i Z_i - \pi(\vec{\mathbf{p}}) \\ &\quad + \lambda\left[\sum_{i=2}^{N} t_i M_i - p_1 g(Z_2,\ldots,Z_N) + \sum_{i=2}^{N} p_i Z_i - B\right] \end{aligned}$$

Optimal Taxation

To derive the Diamond–Mirrlees results, begin by considering the first-order conditions with respect to t_k and computing $\partial L/\partial t_k$ as an intermediate step:

$$\begin{aligned} \frac{\partial L}{\partial t_k} &= M_k + \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} - M_k - \sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k}\right) \\ &\quad - \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial t_k} - \sum_{i=2}^{N} \pi_i \frac{\partial p_i}{\partial t_k} \quad k = 2,\ldots,N \end{aligned} \tag{24.9}$$

2. Recall from the discussion of the optimal tax problem that loss minimization requires measurement at the compensated equilibrium.

3. Refer to Chapter 14 for a discussion of the compensated market clearance relationships.

Multiply each market clearance Eqn (24.6) by $\partial p_i/\partial t_k$ and sum over all $(N-1)$ relationships to obtain

$$\sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} = \sum_{i=2}^{N} \pi_i \frac{\partial p_i}{\partial t_k} + \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial t_k} \tag{24.10}$$

Therefore, Eqn (24.9) simplifies to

$$\frac{\partial L}{\partial t_k} = -\sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) \quad k = 2, \ldots, N \tag{24.11}$$

Next, differentiate the government budget constraint with respect to t_k:

$$\frac{\partial B}{\partial t_k} = M_k + \sum_{i=2}^{N} t_i \left(M_{ik} + \sum_{j=2}^{N} M_{ij} \frac{\partial p_j}{\partial t_k} \right) + \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial t_k} \quad k = 2, \ldots, N \tag{24.12}$$

From Eqn (24.11)

$$\frac{\partial B}{\partial t_k} = -\frac{\partial L}{\partial t_k} + \left(M_k + \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial t_k} \right) \tag{24.13}$$

With CRS in private production, $\sum_{i=1}^{N} \pi_i (\partial p_i/\partial t_k) = 0$. But $\partial p_1/\partial t_k = 0$, so

$$\sum_{i=2}^{N} \pi_i \frac{\partial p_i}{\partial t_k} = 0 \tag{24.14}$$

Hence, from Eqn (24.10)

$$\sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial t_k} = \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} \tag{24.15}$$

Given Eqns (24.15), (24.13) can be rewritten as

$$\frac{\partial B}{\partial t_k} = -\frac{\partial L}{\partial t_k} + \left(M_k + \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} \right) \tag{24.16}$$

Combining Eqns (24.11) and (24.16) and incorporating the Lagrangian multiplier λ, the first-order conditions with respect to t_k are

$$(1-\lambda)\frac{\partial L}{\partial t_k} + \lambda \left(M_k + \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial t_k} \right) = 0 \quad k = 2, \ldots, N \tag{24.17}$$

But Eqn (24.17) is identical to Eqn (14.28), the first-order conditions when revenue was simply raised for its own sake and returned lump sum. Applying the manipulations of Chapter 14, these conditions imply the standard optimal commodity tax rule, Eqn (13.32), or

$$\frac{\sum_{i=2}^{N} t_i M_{ik}}{M_k} = \frac{\lambda}{1-\lambda} = C \quad k = 2, \ldots, N \tag{24.18}$$

Thus, introducing government production into the analysis does not alter the optimal tax rules, the first of the two main Diamond–Mirrlees results. As was noted in Chapter 14, this result depends crucially on the assumption of CRS in private production.

Optimal Government Production

To derive their second, more striking result, differentiate the first-order conditions with respect to the Z_k. As before, begin with a preliminary consideration of $\partial L/\partial Z_k$:

$$\frac{\partial L}{\partial Z_k} = \sum_{i=2}^{N} M_i \frac{\partial p_j}{\partial Z_k} - \sum_{i=2}^{N} \sum_{j=2}^{N} t_i M_{ij} \frac{\partial p_j}{\partial Z_k} + p_1 g_{Z_k} - p_k - \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial Z_k} - \sum_{i=2}^{N} \pi_i \frac{\partial p_i}{\partial Z_k} \quad k = 2, \ldots, N \tag{24.19}$$

Multiplying the market clearance Eqn (24.6) by $\partial p_i/\partial Z_k$ and summing over all $(N-1)$ relationships yield

$$\sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial Z_k} = \sum_{i=2}^{N} \pi_i \frac{\partial p_i}{\partial Z_k} + \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial Z_k} \tag{24.20}$$

Hence, Eqn (24.19) simplifies to

$$\frac{\partial L}{\partial Z_k} = -\sum_{i=2}^{N} \sum_{j=2}^{N} t_i M_{ij} \frac{\partial p_j}{\partial Z_k} + p_1 g_{Z_k} - p_k \quad k = 2, \ldots, N \tag{24.21}$$

Next, consider $\partial B/\partial Z_k$:

$$\frac{\partial B}{\partial Z_k} = \sum_{i=2}^{N} \sum_{j=2}^{N} t_i M_{ij} \frac{\partial p_j}{\partial Z_k} - p_1 g_{Z_k} + p_k + \sum_{i=2}^{N} Z_i \frac{\partial p_i}{\partial Z_k} \quad k = 2, \ldots, N \tag{24.22}$$

From market clearance and CRS in private production, Eqn (24.22) can be restated as

$$\frac{\partial B}{\partial Z_k} = \sum_{i=2}^{N} \sum_{j=2}^{N} t_i M_{ij} \frac{\partial p_j}{\partial Z_k} - p_1 g_{Z_k} + p_k + \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial Z_k} \tag{24.23}$$

Thus,

$$\frac{\partial B}{\partial Z_k} = -\frac{\partial L}{\partial Z_k} + \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial Z_k} \quad k = 2, \ldots, N \tag{24.24}$$

Combining Eqns (24.21) and (24.24) and incorporating λ, the first-order conditions with respect to the Z_k are

$$(1-\lambda)\frac{\partial L}{\partial Z_k} + \lambda \sum_{i=2}^{N} M_i \frac{\partial p_i}{\partial Z_k} \quad k = 2, \ldots, N \tag{24.25}$$

Substituting the expression for $\partial L/\partial Z_k$ and changing the summation index on the $M_i \partial p_i/\partial Z_k$ terms, Eqn (24.25) becomes

$$(1-\lambda)\left[-\sum_{i=2}^{N}\sum_{j=2}^{N} t_i M_{ij}\frac{\partial p_j}{\partial Z_k} - p_k + p_1 g_{Z_k}\right] + \lambda \sum_{j=2}^{N} M_j \frac{\partial p_j}{\partial Z_k} = 0 \quad k = 2,\ldots,N \tag{24.26}$$

To consider the effect that optimal taxation has on these rules, rewrite (24.26) as

$$\sum_{j=2}^{N}\left[(1-\lambda) - \sum_{i=2}^{N} t_i M_{ij} + \lambda M_j\right]\frac{\partial p_j}{\partial Z_k} + (1-\lambda)\left(p_k - p_1 g_{Z_k}\right) = 0 \quad k = 2,\ldots,N \tag{24.27}$$

But, if commodity taxes are set optimally in accordance with Eqn (24.18):

$$-(1-\lambda) - \sum_{i=2}^{N} t_i M_{ij} + \lambda M_j = 0 \quad j = 2,\ldots,N \tag{24.28}$$

Thus, the government production rule is simply

$$p_k - p_1 g_{Z_k} = 0 \quad k = 2,\ldots,N \tag{24.29}$$

or

$$p_k = p_1 g_{Z_k} \quad k = 2,\ldots,N \tag{24.30}$$

Equation (24.30) is the standard first-best rule for production efficiency in competitive markets. Alternatively,

$$\frac{p_k}{p_j} = \frac{g_{Z_k}}{g_{Z_j}} = \mathrm{MRT}_{Z_k,Z_j} \quad k,j = 2,\ldots,N \tag{24.31}$$

with the government using the competitively determined producer prices as shadow prices in its production decisions.

This may well be the most striking result in all of second-best public expenditure theory, one of the precious few examples of a simple second-best decision rule. It implies overall production efficiency for the economy[4] or that the economy should remain on its aggregate production–possibilities frontier. Of course, with distorting taxation, the economy cannot also be on its first-best utility–possibilities frontier. A final implication in an intertemporal context is that government investment decisions should use the private sector's gross-of-tax returns to capital as the rate of discount in present value calculations (recall that p_k is a gross-of-tax price for an input such as capital).[5] Since the U.S. marginal corporate tax rate is 34% or 35% for most firms in the United States, this implies a fairly high government rate of discount, the rate of return the government must beat to justify public investment at the expense of private investment.

PRODUCTION DECISIONS WITH NONOPTIMAL TAXES

The Diamond–Mirrlees problem provides a clear example of just how far removed second-best theory often is from the complexities of the real world, even though it contains elements that are more realistic than the traditional first-best assumptions. Taxes are distorting in this model, but assuming that current tax rates are (even approximately) at their optimal values is every bit as heroic as assuming that taxes are (approximately) lump sum, which first-best theory requires. We can move somewhat closer to reality by assuming explicitly that the current rates are nonoptimal and asking how this affects the government's production decision rules. Formally, this assumption is equivalent to adding further constraints to the original Diamond–Mirrlees problem of the form that a subset of the tax rates is predetermined at nonoptimal levels. Given these predetermined rates, the first-order conditions of the new problem indicate how the government can adjust its production decisions to minimize loss.

Unfortunately, the resulting production rules are extremely complex. They have a plausible interpretation, but it is doubtful whether any government would have sufficient information to implement them. Furthermore, this problem is still far removed from reality, for it retains the assumption of a perfectly competitive CRS private production sector. Were we to introduce monopoly elements in private production and/or decreasing or increasing returns to scale with pure profits or losses, the optimal production rules would change once again. Consequently, the normative policy content of this model is not especially compelling either. Nonetheless, it is instructive to explore the production decision rules when taxes are nonoptimal if only to give a flavor for this kind of analysis.

To keep the notation as simple as possible, rewrite the loss function entirely in vector notation as

$$L(t;Z) = M(q;\overline{U}^0) - t'M_i + p_1 g(Z) - (q-t)'Z - \pi(q-t) \tag{24.32}$$

4. Recall that the private sector is assumed to be perfectly competitive and therefore first-best pareto efficient.

5. Intertemporally, all budget constraints in the general equilibrium framework must balance in terms of present value, not year by year, and there must be perfect capital markets for borrowing and lending.

Written in this form, the loss function incorporates every relevant constraint except for the market clearance equation, Eqn (24.6), expressed in vector notation as

$$M_i\left(q;\vec{\mathbf{U}}^0\right) = \pi_i(q-t)+Z \tag{24.33}$$

The nonoptimal tax and production rules are derived by totally differentiating the loss function with respect to t and Z, and using Eqn (24.33) to simplify the resulting expression:

$$\begin{aligned} dL(t;Z) &= M_i'\frac{\partial q}{\partial t}dt + M_i'\frac{\partial q}{\partial Z}dZ - M_i'dt - t'M_{ij}\frac{\partial q}{\partial t}dt \\ &-t'M_{ij}\frac{\partial q}{\partial Z}dZ + p_1 g_Z dZ - Z'\frac{\partial q}{\partial t}dt + Z'dt - Z'\frac{\partial q}{\partial Z}dZ \\ &-(q-t)'dZ - \pi_i'\frac{\partial q}{\partial t}dt - \pi_i'\frac{\partial q}{\partial Z}dZ + \pi' dt \end{aligned} \tag{24.34}$$

From market clearance:

$$M_i'dt = \pi_i'dt + Z'dt \tag{24.35}$$

$$M_i'\frac{\partial q}{\partial t}dt = \pi_i'\frac{\partial q}{\partial t}dt + Z'\frac{\partial q}{\partial t}dt \tag{24.36}$$

and

$$M_i'\frac{\partial q}{\partial Z}dZ = \pi_i'\frac{\partial q}{\partial Z}dZ + Z'\frac{\partial q}{\partial Z}dZ \tag{24.37}$$

Also,

$$(q-t)' = p' \tag{24.38}$$

Using Eqns (24.35) to (24.38), Eqn (24.34) simplifies to

$$\begin{aligned} dL(t;Z) &= -t'M_{ij}\frac{\partial q}{\partial t}dt - t'M_{ij}\frac{\partial q}{\partial Z}dZ \\ &+ p_1 g_Z dZ - p'dZ \end{aligned} \tag{24.39}$$

Next, totally differentiate Eqn (24.8), obtaining:

$$dq = \frac{\partial q}{\partial t}dt + \frac{\partial q}{\partial Z}dZ \tag{24.40}$$

Substituting Eqn (24.40) into (24.39) yields

$$dL(t;Z) = -t'M_{ij}dq + p_1 g_Z dZ - p'dZ \tag{24.41}$$

The first point to notice is that the Diamond–Mirrlees production rules follow directly from Eqn (24.41). Suppose taxes are set optimally. Since setting taxes is equivalent to setting consumer prices, this means that the vector dq is also optimal. But at the optimum, $dL = 0$. Hence, optimal taxation implies a dq such that the first term in Eqn (24.41) is zero ($dL = -t'M_{ij}dq = -t'dX = 0$ at the optimum, t^*). The vector dZ must also be compatible with $dL = 0$.

Hence,

$$p_1 g_Z dZ - p'dZ = 0 \tag{24.42}$$

or

$$p_1 g_Z = p' \tag{24.43}$$

Tax Rules

If taxes are not optimal, however, the decision rules for government production are more complex, since changes in Z change q, thereby indirectly affecting dL through the (nonzero) tax term in Eqn (24.41). The separate effects of taxes and government production on loss in the general case can be obtained by totally differentiating the market clearance equations, solving for dq in terms of dt and dZ, and substituting the resulting expression for dq into the first term of Eqn (24.41), as follows:

$$M_{ij}dq = \pi_{ij}dq - \pi_{ij}dt + dZ \tag{24.44}$$

$$dq = \left(-\pi_{ij}dt + dZ\right)E^{-1} \tag{24.45}$$

where

$E = [M_{ij} - \pi_{ij}]$, the matrix of compensated demand and private production price derivatives (as defined in Chapter 14).

Substituting Eqn (24.45) into (24.41) and rearranging terms:

$$dL = t'\left(M_{ij}\right)E^{-1}\pi_{ij}dt - \left(-p_1 g_Z + p' + t'M_{ij}E^{-1}\right)dZ \tag{24.46}$$

Equation (24.46) can be used to compute the change in loss, or welfare, resulting from any combination of changes in t and Z, with the remaining t and Z held constant. One immediate and important implication of Eqn (24.46) is that the addition of government production does not affect any of the theorems in Chapter 14 on the deadweight loss from changes in tax rates. In this sense, the welfare effects of government production are separable from those of distorting taxes. The term $[t'M_{ij}E^{-1}\pi_{ij}]dt$ is identical to the right-hand side (RHS) of Eqn (14.17) in Chapter 14, with:

$$M_{ij} = \frac{\partial X^{comp}}{\partial q} \quad \text{and} \quad \pi_{ij} = \frac{\partial Y}{\partial p}$$

As indicated in Chapter 14, the marginal loss from a small increase in a distorting tax can be interpreted as a change in consumer and producer surpluses, where consumer surplus is defined in terms of compensated demand curves. This result continues to hold in the presence of

government production because with $dZ = 0$, the market clearance derivatives imply $dq = -\pi_{ij}E^{-1}dt$, or

$$dL(t;Z) = t'M_{ij}dq = t'\frac{\partial X}{\partial q}dq = t'dX = (q-p)'dX \tag{24.47}$$

exactly as in Chapter 14.

Production Rules

The government's production rules can be stated in a number of different ways depending upon the manner in which the control variables are manipulated. The most straightforward example to consider is the welfare implication of marginally increasing one of the inputs, say Z_k, in order to increase output of Z_1 through the marginal product relationship, g_{Z_k}, all other Z and the tax rates being constant. According to Eqn (24.46), the change in loss from this move would be

$$dL(\bar{t};Z) = -(-p_1 g_{Z_k} + p_k + t'M_{ik}E^{-1})dZ_k \tag{24.48}$$

The *optimal* adjustment of Z_k is one for which $dL = 0$, or

$$-(-p_1 g_{Z_k} + p_k + t'M_{ik}E^{-1})dZ_k = 0 \tag{24.49}$$

In a first-best environment, the government would hire Z_k until its price equaled the value of its marginal product, or $p_1 g_z = p_k$ (recall that Z_k enters negatively in $g(z)$). With distorting and nonoptimal taxes, however, Eqn (24.49) implies that the true social costs of hiring Z_k are $(p_k + t'M_{ik}E^{-1})$. Hence, the government should use these true costs as the shadow price for decision-making and equate them to the value of marginal product. In other words, set

$$p_1 g_z = (p' + t'M_{ij}E^{-1}) \tag{24.50}$$

The term $t'M_{ij}E^{-1}$ turns out to have an intuitively appealing interpretation. With taxes held constant, $dt = 0$, $dq = dp$, and the market clearance derivatives, Eqn (24.44), become

$$M_{ij}dq = \pi_{ij}dp + dZ \tag{24.51}$$

Substituting $dq = dp$ and solving for dZ yield

$$dZ = (m_{ij} - \pi_{ij})dq = Edq \tag{24.52}$$

Substituting Eqn (24.52) into the last term of Eqn (24.46) and letting only Z_k change, we obtain

$$-(-p_1 g_{Z_k}dZ_k + p_k dZ_k + t'M_{ik}dq) = 0 \tag{24.53}$$

Rearranging terms:

$$p_1 g_{Z_k} = \left(p_k + t'M_{ik}\frac{dq}{dZ_k}\right) \tag{24.54}$$

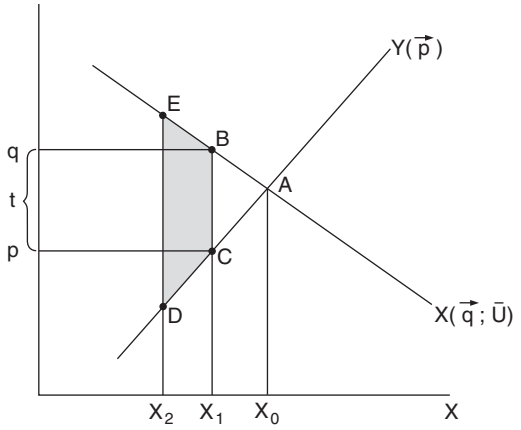

FIGURE 24.1

But,

$$t'M_{ik}\frac{dq}{dZ_k} = t'\frac{\partial X}{\partial q}\frac{dq}{dZ_k} = t'dX \tag{24.55}$$

the change in tax revenues caused by a change in Z_k at constant tax rates. This revenue change represents an additional deadweight burden to the consumer because, with nonzero taxes, changes in market equilibria change the sum of producers' and consumers' surpluses lost as a result of the tax distortions.

Fig. 24.1 illustrates this point. X_0 is the original no-tax equilibrium, and X_1 is the equilibrium with taxes (and government production), with loss equal to triangle ABC. If a marginal increase in Z_k shifts X again, generating a new equilibrium, X_2, then the loss area in the market for X increases by the trapezoidal area EBCD, which is approximately equal to $t'dX$ for small changes. The full social opportunity costs of using Z_k, then, are the standard market opportunity costs, p_k, plus the additional excess burden implied by the tax revenue response to changes in Z_k. Finally, since this result holds for all k, the production rule can be expressed in the traditional format as $\text{MRT}_{i,j} = \theta_i/\theta_j$, for $j = 2, \ldots, N$, where $\theta_i = p_i + t'M_{ji}(dq/dZ_i)$, the optimal shadow price for good (factor) i.

Special Cases

The government production shadow prices $\theta_k = (p_k + t'M_{ik}E^{-1})$ have a very appealing weighted-average interpretation if one assumes, as an approximation, that all cross-price derivatives in demand and private production are zero.[6] With this assumption, the shadow price simplifies to

$$\theta_k = p_k + t_k\frac{\partial X_k}{\partial q_k}\left[\frac{\partial X_k}{\partial q_k} - \frac{\partial Y_k}{\partial p_k}\right]^{-1} \tag{24.56}$$

6. This assumption is not tenable for either the compensated demand derivatives or the production derivatives. At least one M_{ij}, for $i \neq j$, and one π_{ij}, for $i \neq j$, must be positive. Thus, the assumption can only be approximately true. See Boadway (1975).

$$\theta_k = p_k + (q_k - p_k)\frac{\partial X_k}{\partial q_k}\left[\frac{\partial X_k}{\partial q_k} - \frac{\partial Y_k}{\partial p_k}\right]^{-1} \tag{24.57}$$

Rearranging the second term on the RHS of Eqn (24.57) yields

$$\theta_k = p_k + (q_k - p_k)\left[\frac{1}{1 - \frac{\frac{\partial Y_k}{\partial p_k}}{\frac{\partial X_k}{\partial q_k}}}\right] \tag{24.58}$$

Let

$$\alpha = \frac{\frac{\partial Y_k}{\partial p_k}}{\frac{\partial X_k}{\partial q_k}}$$

Therefore

$$\theta_k = p_k + \left(q_k - p_k\right)\left[\frac{1}{1-\alpha}\right] \tag{24.59}$$

Rearranging terms:

$$\theta_k = q_k\left(\frac{1}{1-\alpha}\right) + p_k\left(\frac{-\alpha}{1-\alpha}\right) \tag{24.60}$$

Equation (24.60) says that the optimal shadow price for the input Z_k is a weighted average of the consumer and producer prices, with the weights equal to the proportions in which the increased Z_k comes at the expense of either decreased demand for the input by the private sector or increased supply of the input from consumers. Given market clearance, these are the only possibilities.

Consider the extreme cases as an aid to intuition. If, on the one hand, the entire increase in Z_k comes from an increase in consumer supply, $\partial Y_k/\partial p_k = 0$, $\alpha = 0$, and $\theta_k = q_k$. The only opportunity cost of increasing Z_k is the private opportunity cost to the consumer of supplying the additional Z_k. If, on the other hand, the entire increase in Z_k comes from a decrease in private demand, $\partial X_k/\partial q_k = 0$, $\alpha \to \infty$, and $\theta_k = p_k$, the market opportunity cost for Z_k. This is effectively what happens with optimal taxation—all changes in government production come entirely at the expense of private production.

The case of $\partial X_k/\partial q_k = 0$ also applies for linear technologies with fixed producer prices. All changes in government production must come entirely at the expense of private production when private production input demands and output supplies are perfectly elastic at the fixed producer prices. Indeed, we would expect the optimal shadow prices to equal the producer prices, $\vec{\mathbf{p}}$ even if cross-price derivatives are nonzero. This can be seen directly from Eqn (24.41). With $\vec{\mathbf{p}}$ constant, $dq = dt$ and Eqn (24.41) becomes

$$dL = -t'M_{ij}dt + \left(p_1 g_z - p'\right)dZ \tag{24.61}$$

Even with nonoptimal distorting taxes, then, the government should use the competitive private sector producer prices as shadow prices to avoid any additional increases in deadweight loss.

Balanced-Budget Changes in *t* and *Z*

Thus far, government production variables have been allowed to change without any reference to the government's budget constraint. Any changes in the budget surplus (deficit) are simply returned to the consumer lump sum. If, in fact, the government is required to maintain budgetary balance, then increasing Z_k may require a simultaneous change in at least one of the tax rates. One can imagine the following policy: Suppose the government increases Z_k (and, implicitly, Z_1) and simultaneously changes the jth tax, t_j, to maintain a balanced budget. Under these circumstances, what is the appropriate shadow price for Z_k?

The solution is straightforward given Eqn (24.46). Totally differentiate the government's budget, $t'M_i - p_1 g(Z) + p'Z = \mathrm{B}$, with respect to t_j and Z_k to determine the required change in t_j for any given (small) change in Z_k, and substitute the resulting solution $dt_j^* = f' \cdot dZ_k$ into Eqn (24.46) to obtain an expression for the change in loss solely as a function of dZ_k. The optimal shadow price is then computed by setting $dL = 0$. Without actually carrying out the calculations, the effect of the budget constraint on the optimal shadow prices can be seen from Eqn (24.46):

$$\frac{dL}{dZ_k} = t'M_{ij}E^{-1}\pi_{ij}\frac{dt_j^*}{dZ_k} - \left(-p_1 g_{Z_k} + p_k + t'M_{ik}E^{-1}\right) = 0 \tag{24.62}$$

where

$\frac{dt_j^*}{dZ_k}$ is the required change in t_j for maintaining budgetary balance.

Note that $\frac{dt_j*}{dZ_k}$ is not the only possibility; any number of tax changes could be used to keep the budget in balance as Z_k changes. The analysis must specify exactly how the tax rates are being changed. In any event, the new shadow price to be equated to the value of marginal product $p_1 g_{Z_k}$ is

$$\theta_k = p_k + t'M_{ik}E^{-1} - t'M_{ij}E^{-1}\pi_{ij}\frac{dt_j^*}{dZ_k} = p_1 g_{Z_k} \tag{24.63}$$

There are now two necessary adjustments to p_k, the private opportunity costs, to obtain the full social opportunity costs of Z_k. The first is the additional deadweight loss as tax revenues adjust directly to the change in Z_k, measured at constant tax rates, the effect described above. The second is the additional deadweight loss resulting from the required increase in t_j to maintain budgetary balance. Since marginal changes in different Zs affect the budget

equation differently, this second source of additional burden is, in general, unique to each Z. Quite obviously, governments are going to have a most difficult time computing these optimal shadow taxes, unless the distorting taxes are optimal or technology is linear. As we have seen, in either of these cases the optimal shadow prices are just the p_k. Note also that using p_k implies that the government's budget constraint necessarily holds. If $p_1 g_{Z_k} = p_k$, then

$$-p_1 g_{Z_k} dZ_k + p_k dZ_k = 0 \quad k = 2, \ldots, N \tag{24.64}$$

implies

$$-p_1 dZ_1 + p_k dZ_k = 0 \quad k = 2, \ldots, N \tag{24.65}$$

SECOND-BEST PRODUCTION RULES WHEN EQUITY MATTERS

Assuming a one-consumer-equivalent economy in second-best analysis is always somewhat contradictory. Unless consumers' tastes are severely restricted, one-consumer equivalence implies that the government is optimally redistributing income lump sum in accordance with the first-best interpersonal equity conditions, thereby equilibrating social marginal utilities of income. But if the government can do this, why would it ever have to use distorting taxes?

The more natural approach in a second-best framework is to deny the existence of optimal income redistribution and assume explicitly that social marginal utilities of income are unequal. This means, however, that the optimal shadow prices for government production decisions depend upon both efficiency and equity considerations, just as the many-person optimal tax and nonexclusive goods decision rules were seen to incorporate both efficiency and equity terms. This is doubly discouraging for policy purposes, as we noted when discussing those problems. Not only are optimal prices further complicated by the addition of equity terms, but also society may not agree on the proper equity weights for each individual. Thus, the analysis runs the risk of becoming totally subjective, since different sets of ethical weights imply different optimal shadow prices. Nonetheless, if society can agree on a ranking of social marginal utilities of income (a big if), then the proper shadow prices for government production can be determined. Furthermore, the shadow prices can be expressed as a simple combination of distinct equity and efficiency effects, at least for the particular government production decisions and second-best distortions being considered in this chapter.

To relate the many-person results as closely as possible to the one-person rules, we will assume away all sources of lump-sum income by requiring that all factor supplies are variable and private production exhibits CRS. A further assumption is that the government budget exactly balances. These assumptions greatly simplify the analysis, while capturing the flavor of many-person second-best analysis.[7]

The government's objective function, then, is

$$W = W\left[V^h\left(\vec{\mathbf{q}}\right)\right] = V\left(\vec{\mathbf{q}}\right) \tag{24.66}$$

where W is the agreed-upon individualistic Bergson–Samuelson social welfare function whose arguments are the individuals' indirect utility functions $V^h(\vec{\mathbf{q}})$. Differentiating totally,

$$dW = \sum_{h=1}^{H} \sum_{i=1}^{N} \frac{\partial W}{\partial V^h} \frac{\partial V^h}{\partial q_i} dq_i = -\sum_{h=1}^{H} \sum_{i=1}^{N} \beta^h X_{hi} dq_i \tag{24.67}$$

from Roy's Identity and the definition of an individual's social marginal utility of income as $\beta^h = (\partial W/\partial V^h)\alpha^h$, where α^h is the private marginal utility of income of person h. It will be convenient to express the change in social welfare in terms of Martin Feldstein's distributional coefficient for X_i:

$$R_i = \sum_{h=1}^{H} \beta^h \frac{X_{hi}}{X_i} \quad i = 1, \ldots, N \tag{24.68}$$

to work with aggregate consumption.[8] R_i is a weighted average of the individuals' social marginal utilities of income, with the weights equal to the proportion of good (factor) i consumed (supplied) by person h. Substituting Eqn (24.68) into (24.67) yields

$$dW = -\sum_{i=1}^{N} R_i X_i dq_i \tag{24.69}$$

The problem is to define dW in terms of the government control variables $\vec{\mathbf{t}} = (t_2, \ldots, t_N)$ and $\vec{\mathbf{Z}} = (Z_2, \ldots, Z_N)$, given the following constraints:

1. Private production possibilities, $F(Y_1, \ldots, Y_N) = 0$, assumed to exhibit CRS.
2. The government production function, $G(Z_1, \ldots, Z_N) = 0$, or $Z_1 = -g(Z_2, \ldots, Z_N)$, with inputs measured negatively.
3. The government budget constraint, $\sum_{i=2}^{N} t_i X_i + \sum_{i=1}^{N} p_i Z_i = 0$.
4. N market clearance relationships, $X_i(\vec{\mathbf{q}}) = Y_i(\vec{\mathbf{p}}) + Z_i$, for $i = 1, \ldots, N$. All markets clear in the actual general equilibrium.
5. $\vec{\mathbf{q}} = \vec{\mathbf{p}} + \vec{\mathbf{t}}$, with $q_1 \equiv p_1 \equiv 1$ and $t_1 \equiv 0$.

7. With only minor changes, the analysis of this section is taken directly from Boadway (1976).
8. Feldstein (1972). Also see our discussion of Feldstein's distributional coefficient in Chapter 14.

As always, the first good serves as the untaxed numeraire.

The analysis proceeds much as in the one-consumer case. Begin by totally differentiating the market clearance equations:

$$dX_i = dY_i + dZ_i \quad i = 1, \ldots, N \tag{24.70}$$

Multiply each equation by $q_i = (p_i + t_i)$ and sum over all N equations to obtain

$$\sum_{i=1}^{N} q_i dX_i = \sum_{i=1}^{N} (p_i + t_i) dY_i + \sum_{i=1}^{N} (p_i + t_i) dZ_i \tag{24.71}$$

Equation (24.71) can be simplified as follows. Totally differentiate the individual consumers' budget constraints $\sum_{i=1}^{N} q_i X_{hi} = 0$, for $h = 1, \ldots, H$, and sum over all individuals to obtain

$$\sum_{i=1}^{N} q_i dX_i = -\sum_{i=1}^{N} X_i dq_i \tag{24.72}$$

Next, differentiate the aggregate private production possibilities $F(\vec{\mathbf{Y}}) = 0$,

$$\sum F_i dY_i = 0 \tag{24.73}$$

But, if markets are perfectly competitive,

$$\frac{F_i}{F_1} = \frac{p_i}{p_1} = p_i, \quad \text{with} \quad p_1 \equiv 1 \quad i = 2, \ldots, N \tag{24.74}$$

Therefore

$$\sum_{i=1}^{N} F_i dY_i = 0 = F_1 \sum_{i=1}^{N} p_i dY_i \tag{24.75}$$

or

$$\sum_{i=1}^{N} p_i dY_i = 0 \tag{24.76}$$

Substituting Eqns (24.72) and (24.76) into (24.71) yields

$$-\sum_{i=1}^{N} X_i dq_i = \sum_{i=1}^{N} t_i dY_i + \sum_{i=1}^{N} t_i dZ_i + \sum_{i=1}^{N} p_i dZ_i \tag{24.77}$$

Using Eqn (24.70), Eqn (24.77) can be expressed as

$$-\sum_{i=1}^{N} X_i dq_i = \sum_{i=1}^{N} t_i dX_i + \sum_{i=1}^{N} p_i dZ_i \tag{24.78}$$

Substituting Eqn (24.78) into (24.69) yields

$$dW = -\sum_{i=1}^{N} R_i X_i dq_i + \sum_{i=1}^{N} X_i dq_i + \sum_{i=1}^{N} t_i dX_i + \sum_{i=1}^{N} p_i dZ_i \tag{24.79}$$

or

$$dW = \sum_{i=1}^{N} (1 - R_i) X_i dq_i + \sum_{i=1}^{N} t_i dX_i + \sum_{i=1}^{N} p_i dZ_i \tag{24.80}$$

Next, incorporate the government production function, $Z_1 = -g\ (Z_2, \ldots, Z_N)$, and note that $t_1 \equiv 0$; $dq_1 = 0$, to rewrite Eqn (24.80) as

$$\begin{aligned} dW = &\sum_{i=2}^{N} (1 - R_i) X_i dq_i + \sum_{i=2}^{N} t_i dX_i \\ &+ \sum_{i=2}^{N} \left(- p_1 g_{Z_i} + p_i \right) dZ_i \end{aligned} \tag{24.81}$$

To eliminate the dX_i, totally differentiate the individual demand (factor supply) functions $X_{hi} = X_{hi}(\vec{\mathbf{q}})$ $h = 1, \ldots, H$, and sum over all individuals to obtain

$$dX_i = \sum_{j=2}^{N} \frac{\partial X_i}{\partial q_j} dq_j \quad i = 1, \ldots, N \tag{24.82}$$

Substituting Eqn (24.82) into (24.81) and rearranging the terms yield

$$\begin{aligned} dW = &\sum_{i=2}^{N} \sum_{j=2}^{N} \left[(1 - R_i) X_i + t_j \frac{\partial X_j}{\partial q_i} \right] dq_i \\ &+ \sum_{i=2}^{N} \left(- p_1 g_{Z_i} + p_i \right) dZ_i \end{aligned} \tag{24.83}$$

Finally, use the market clearance equations:

$$X_i(\vec{\mathbf{q}}) = Y_i(\vec{\mathbf{q}} - \vec{\mathbf{t}}) + Z_i \quad i = 1, \ldots, N \tag{24.84}$$

to express dq_i in terms of the control variables dt_i and dZ_i, as follows. From Walras' law, only $(N-1)$ of these relationships are independent. Since good 1 is the numeraire, eliminate the first equation and totally differentiate Eqn (24.2), for $i = 2, \ldots, N$ to obtain

$$\begin{aligned} \sum_{j=2}^{N} \frac{\partial X_i}{\partial q_j} dq_j = &\sum_{j=2}^{N} \frac{\partial Y_i}{\partial p_j} dq_j - \sum_{j=2}^{N} \frac{\partial Y_i}{\partial p_j} dt_j + dZ_i \\ &i = 2, \ldots, N \end{aligned} \tag{24.85}$$

Writing all $(N-1)$ equations in matrix notation:

$$\left(\frac{\partial X}{\partial q} \right) dq = \left(\frac{\partial Y}{\partial p} \right) dq - \left(\frac{\partial Y}{\partial p} \right) dt + dZ \tag{24.86}$$

All matrices have dimension $(N-1) \times (N-1)$; all vectors have dimension $(N-1) \times 1$. Solving Eqn (24.86) for dq yields

$$dq = E^{-1} \left[- dt' \left(\frac{\partial Y}{\partial p} \right) + dZ \right] \tag{24.87}$$

where

$$E = \left[\left(\frac{\partial X}{\partial q}\right) - \left(\frac{\partial Y}{\partial p}\right)\right]$$

Substituting Eqn (24.87) into (24.83), rearranging terms, and writing the resulting equation in matrix notation yield

$$dW = -\left[\left[\left[(1-R)'\cdot X\right] + t'\frac{\partial X}{\partial q}\right]E^{-1}\left(\frac{\partial Y}{\partial p}\right)\right]dt + \left[\left[(1-R)'\cdot X\right]E^{-1} + t'\frac{\partial X}{\partial q}E^{-1} - p_1 g_Z + p'\right]dZ \tag{24.88}$$

Equation (24.88) gives the change in social welfare for any given (marginal) changes in the government control variables, evaluated at the existing levels of each t and Z.

Notice that if the distribution of income were optimal, so that $\beta^h = \beta$, for $h = 1, \ldots, H$, then $R_i = R_i = \sum_{h=1}^{N}\beta_h X_{hi}/X_i = \beta$, for $i = 1, \ldots, N$, the common social marginal utility of income. Since W can be defined such that $\beta = 1$, by setting $\partial W/\partial V^h = 1/\alpha^h$, for $h = 1, \ldots, H$, dW simplifies to

$$dW|_{\beta^h=\beta=1} = -\left(t'\frac{\partial X}{\partial q}\right)E^{-1}\left(\frac{\partial Y}{\partial p}\right)dt + \left[t'\left(\frac{\partial X}{\partial q}\right)E^{-1} - p_1 g_Z + p'\right]dZ \tag{24.89}$$

But Eqn (24.89) is identical to Eqn (24.46), with $dW|_{\beta^h=\beta=1} = -dL$, $(\partial X/\partial q) = M_{ij}$, and $(\partial Y/\partial p) = \pi_{ij}$ (i.e., assuming away income effects so that actual and compensated demands and factor supplies are identical). Since Eqn (24.46) captures all the efficiency implications of distorting taxation, Eqn (24.88) can be viewed as a simple linear combination of the efficiency and equity effects of tax distortion, where the latter are embodied in the coefficients $[(1-R)'\cdot X]\ E^{-1}\ (\partial Y/\partial p)$ for changes in tax rates, and $[(1-R)'\cdot X]\ E^{-1}$ for changes in the government production variables. Thus, if the government were able to estimate the efficiency effects of the tax distortions and if it could provide an acceptable set of social marginal utilities of income, adjusting tax and production decision rules for equity considerations would be a relatively straightforward exercise. These are two huge ifs, however. Although it is appealing to be able to separate the equity and efficiency effects of government policies in principle, there is still no reason to suppose that a society will be able to agree on a set of distributional coefficients, R_i, much less that the government can compute the efficiency distortions with any confidence.

To make matters worse, the full social costs (benefits) for public sector inputs and outputs in a many-person environment generally consist of the coefficients on (some of) the dt terms as well as the coefficients on the appropriate dZ terms. To see why, suppose that the government increases its purchase of input Z_k, thereby increasing production of Z_1 through g_{Z_k}. It is tempting to conclude that the social cost for Z_k is p_k plus the appropriate terms in $[[(1-R)'\cdot X]\ E^{-1} + t'(\partial X/\partial q)\cdot E^{-1}]$, to be equated to $p_1 g_{Z_k}$, the value of marginal product for Z_k. But, this ignores the fact that the government's budget must remain in balance. When computing the dq as functions of dt and dZ in Eqn (24.86), we invoked Walras' law to eliminate the market clearance equation for good 1. But the N market clearance equations are dependent only if all consumers are on their budget constraints, all firms are maximizing profits, and the government budget always remains in balance. Thus, although the government budget constraint was never explicitly mentioned in deriving the expression for dW, the solution for dq in Eqn (24.86) and for dX_i in Eqn (24.82) implicitly assumed that it holds, since lump-sum income was held constant. Therefore, any policy experiments evaluated with Eqn (24.88) must be consistent with maintaining the government's budget constraint. Were the government to follow the optimal shadow price for Z_k derived above, the budget will surely not remain in balance, since this would require $p_1 g_{Z_k} = p_k$, or $p_1 dZ_1 = p_k dZ_k$. In general, then, the government must vary at least one of the tax rates to maintain budgetary balance, in which case the full social cost of Z_k contains terms of the form:

$$\left[\left[(1-R)'\cdot X\right] + t'\frac{\partial X}{\partial q}\right]E^{-1}\left(\frac{\partial Y}{\partial p}\right)\frac{dt*}{dZ_k}$$

As before, the dt^*/dZ_k are the tax changes necessary to keep[9]:

$$\sum_{i=2}^{N} t_i X_i + \sum_{i=1}^{N} p_i Z_i = 0$$

The only simple case for optimal shadow prices occurs when the producer prices, $\vec{\mathbf{p}}$, are fixed, such as with linear technologies or for a small country facing perfectly elastic supplies (input demand) at world prices. With $dq = dt$, Eqn (24.83) becomes

$$dW|_{p=\bar{p}} = \sum_{i=2}^{N}\sum_{j=2}^{N}\left[(1-R_i)X_i + t_j\frac{\partial X_j}{\partial q_i}\right]dt_i + \sum_{i=2}^{N}\left(-p_1 g_{Z_i} + p_k\right)dZ_i \tag{24.90}$$

The optimal shadow prices are just the private sector or producer prices p_k, exactly as in the one-consumer

9. The budget could remain in balance without changing taxes if many Zs change simultaneously, but budgetary balance is unlikely to be maintained without changing some taxes.

(equivalent) economy. With perfectly elastic supplies (input demands), changes in government production variables do not change consumer prices. Therefore, they have no equity effects and no efficiency implications other than the requirement that government producers do just as well as the private opportunity costs reflected in the price p_k. Furthermore, as noted in the preceding sections, with $p_1 g_{Z_k} = p_k$, or $p_1 dZ_1 = p_k dZ_k$, for $k = 2, \ldots, N$, marginal changes in government production are always self-financing, so that the government's budgetary balance is automatically maintained.

CONCLUDING COMMENTS

Equation (24.88) provides a fairly comprehensive guideline for government decision-making. The many-person problem considered above imposes no restrictions on the form of the government production function and allows the government to tax all goods and factors. Furthermore, the tax incidence analysis of Chapter 14 showed that it makes no difference whether distorting per-unit taxes are levied on producers or consumers. Finally, Eqn (24.88) holds at the existing values of all government tax and production control variables.

It is important to realize, however, that the analysis is not fully general. There are, for example, no externalities arising from the government's activity, private production is assumed to be perfectly competitive and exhibit CRS, and the government is free to vary all price−cost margins. Changes in any or all of these assumptions can be expected to alter the implied optimal shadow prices for public production.

REFERENCES

Boadway, R., July 1975. Cost−benefit rules and general equilibrium. Review of Economic Studies 42 (3), 361−374.

Boadway, R., November 1976. Integrating equity and efficiency in applied welfare economics. Quarterly Journal of Economics 90 (4), 541−556.

Diamond, P.A., Mirrlees, J., March−June 1971. Optimal taxation and public production (2 parts; Part I: Production Efficiency, Part II: Tax Rules). American Economic Review, 61 (1), 8−27; 61 (3:1), 261−278.

Feldstein, M., March 1972. Distributional equity and the optimal structure of public prices. American Economic Review 62 (1), 32−36.

Chapter 25

Behavioral Public Sector Economics

Chapter Outline

An assumption of long standing in economic analysis is that individuals' preferences are a given. Starting around 1970, economists began consulting with psychologists and psychiatrists and reading the psychological literature to try to understand how preferences are formed. The idea was that achieving a better understanding of the psychological foundations of consumer behavior might lead to better assumptions about how individuals behave, at least in certain circumstances. This line of inquiry has produced dramatic results, to say the least. It spawned a new branch of economics, called behavioral economics, that has posed a fundamental challenge to standard microeconomic theory and therefore, by extension, to the mainstream economic theory of the public sector presented in this textbook. Behavioral economists have called into question nearly all of the standard assumptions underlying the theory of utility maximization by rational, self-interested individuals. And their challenge is gaining support rather than receding. Their message that good psychology makes for good economics has taken hold.

The primary means of exploring the psychological foundations of behavior has been laboratory experiments, such as the experiments described in Chapter 6 to determine if people will tend to free ride in the provision of nonexclusive goods. Typically, the subjects are presented with a situation and asked to make a decision or presented with alternatives and asked to express a preference for one of them. In each instance, standard economic theory indicates what the decision or preference should be. In the nonexclusive goods experiment, the subjects are expected to free ride. If they happen to do something else, and often they do, then the economists and/or psychologists conducting the experiments search for a psychological explanation of what appears to be irrational or anomalous behavior from the standard economic perspective. As we saw in the nonexclusive goods experiment, the subjects almost never chose to entirely free ride, at least not at first, and we described a number of psychologically based explanations as to why this might be.

The results of these experiments are by no means the only source of evidence of anomalous behavior. Behavioral economists can point to many instances in actual situations in which individuals' behavior is not in accordance with standard economic theory.

THE BEHAVIORAL ANOMALIES

The idea that some people behave at times in ways that appear to be contrary to their own best interests is hardly noteworthy in and of itself. There are a lot of people in the world and some will occasionally behave in highly unusual ways. What has made behavioral economics so compelling, however, is that it has uncovered anomalies that are widespread, systematic, realistic, and important. They do not appear to be outliers that can simply be ignored.

Behavioral economists have uncovered an uncomfortably large number of such anomalies. What follows is a partial list of the more important anomalies that are

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00025-X

potentially highly relevant to public sector theory and policy.

PROSPECT THEORY: THE REJECTION OF EXPECTED UTILITY MAXIMIZATION

Psychologists Daniel Kahneman and Amos Tversky were among the first researches to call into question the assumptions of standard economic theory. They conducted a long series of experiments on how people behave in situations involving risk.

Standard theory predicts that people maximize expected utility in risky situations, using Bayes' law to update their priors regarding the probabilities of the various possible states of nature occurring as they receive new information about how frequently each state has occurred. Kahneman and Tversky (KT) found instead that their subjects consistently violated all the assumptions of the standard model. They developed an entirely new model of behavior under uncertainty that they called prospect theory. According to Nicholas Barberis, prospect theory is by far "…the best available description of how people evaluate risk in experimental settings."[1]

Prospect theory contains a number of elements that have become central tenets of behavioral economics and we will discuss it at length in the next section of the chapter. Kahneman received the Nobel Memorial Prize in Economics in 2002 for developing prospect theory and for being in effect the father of behavioral economics.

PRESENT-BIASED PREFERENCES: SELF-CONTROL ISSUES

The standard assumption in economics is that agents compare present and future outcomes by using exponential discounting to compute the present value of all future outcomes, which places all the outcomes on the same basis for comparison. Behavioral economists have found, in contrast, that people have present-biased preferences. They consistently give too much weight to current outcomes over future outcomes relative to exponential discounting of the future. This is also referred to as a self-control problem, in the sense that people want instant gratification; they do not have the self-control to wait for better options in the future.

The following experimental result is an example frequently noted in the literature. Subjects are given the choice of earning \$7.00 per hour one day and resting the next day or resting the first day and earning \$7.70 per hour, 10% more, the next day. If the two days are today and tomorrow, subjects often choose \$7.00 today and rest tomorrow. If the two days are a month from now or later, they choose the option to rest the first day and earn \$7.70 the second day. Under exponential discounting, subjects would always choose the second option since the return for waiting one day—10%—is much larger than any measured, or even conceivable, one-day discount factor. Therefore, choosing the first option appears to be an anomaly, an instance of present bias relative to the rational decision. Moreover it is a rolling bias, repeating itself each period as the future becomes the present. The self-control problem never goes away.

Behavioral economists have proposed that present-biased preferences can be explained by assuming that the people use what David Laibson called quasi-hyperbolic discounting of the future rather than exponential discounting. Under quasi-hyperbolic discounting, utility over time is evaluated at any time t as

$$U^t(U_t, U_{t+1}, \ldots, U_T) = \delta^t U_t + \beta \sum_{\lambda=t+1}^{T} \delta^\lambda U_\lambda \qquad (25.1)$$

where δ = the standard exponential discount factor $\left(\frac{1}{1+r_t}\right)$ *or* $\left(\frac{1}{1+r_\lambda}\right)$ and $0 < \beta < 1$. $\beta = 1$ is the standard exponential discounting, so that $\beta < 1$ captures the present bias.

Return to the example above of choosing when to rest and when to work. Suppose δ is close to one, as it would be in a one-day time frame, and $\beta = 0.8$. If the decision of when to rest and work is made today, then the \$7.70 earned tomorrow has to be multiplied by 0.8 when comparing it with the \$7.00 that can be earned today. 0.8 (\$7.70) = \$6.16. Taking \$7.00 today is the better option. But the 0.8 factor applies to both options one month from now, in which case taking \$7.70 the next day is the better option under any reasonable discount factor δ.[2]

One does not have to resort to experiments to uncover self-control problems. Heavy drinkers and smokers often admit that they would be better off in the long run if they stopped drinking and smoking but they lack the self-control to stop. As behavioral economist Matthew Rabin has said: "Common sense, millennia of folk wisdom, and hundreds of psychological experiments all support present-biased preferences."[3]

SOCIAL PREFERENCES

Economists of all persuasions have long accepted that people are not entirely self-interested in their economics affairs. We noted in Chapter 10 that public choice economists explain the existence of public transfer programs as

1. Barberis (2012), p. 2. Kahneman and Tversky's seminal article was Kahneman and Tversky (1979).

2. The example appears in Rabin (2002). The quotation is on p. 669. Matthew Rabin is one of the leading behavioral economists and his Alfred Marshal Lecture is an excellent tour of the principal behavioral anomalies Laibson (1994).

3. Rabin, *op. cit.*, p. 669.

just another example of a consumer externality. The existence of private charity indicates that the nonpoor are at least somewhat altruistic toward the poor: Something about the poor bothers them, either they lack a proper amount of food, or housing, or medical care, or simply lack the resources to enjoy even a minimally acceptable standard of living. Charity is driven into the public sector because each nonpoor person has an incentive to free ride on the altruistic or charitable impulses of all the other nonpoor.

Behavioral economists have significantly refined the nature of these other-directed or social preferences. The altruism discussed in Chapter 10 is referred to as pure altruism. As noted in the chapter, James Andreoni showed that the implications of private charitable giving under pure altruism are not borne out at all in the United States. For example, the model implies that only the richest individuals will give privately; all the others will free ride. In fact, charitable giving is widespread in the United States.

Ernst Fehr has been one of the leading researchers among behavioral economists on social preferences, using experimental techniques. His experiments, undertaken with various coauthors, show that people exhibit what he calls reciprocal altruism rather than pure altruism. Subjects care about the character revealed by the other subjects in experimental situations and respond in-kind. They are both conditional cooperators and willing punishers. An experiment frequently cited as an example is the ultimatum or dictator game. One subject, the dictator, proposes a split of $100 between himself and another subject. Then the second subject can either accept or reject the proposed split. If the second subject is entirely rational in terms of the standard economic model, then he should be willing to accept just a penny, because anything is better than nothing. Yet if the proposed split wanders too far from 50/50, then the proposal is often rejected. The second subject is willing to pass up his gain and punish the dictator for failing to propose what he views as a fair or reasonable split. But he will accept a less-than-even split if it is fairly close to 50/50; that is viewed as reasonable and turns him into a conditional cooperator. Similarly, in the public goods experiments, subjects are more willing to play the public good as the rounds continue if they see others contributing to the public good—they are again conditional cooperators.[4]

Moving beyond the experiments, behavior economists believe that reciprocal altruism is the explanation for people's willingness to pay taxes. The Allingham/Sandmo model of tax evasion described in Chapter 15 makes the standard assumption that taxpayers are entirely self-interested. They are quite willing to exploit private information about their incomes and evade taxes. Therefore, the government has to dissuade them from evading by auditing returns and/or imposing stiff penalties on evaders. Yet most researchers have concluded that tax evasion is much less in the industrial market economies than would be predicted by the Allingham/Sandmo model, given actual auditing rates and penalties. The empirical research has shown that tax evasion is not just a matter of self-interest. It is also related to people's view of their governments along such dimensions as the perceived fairness of the tax code, trust in the government to do the right thing, their evaluation of the usefulness of government expenditures, and the level of political corruption. According to the behavioral perspective, taxpayers exhibit reciprocal altruism with the government as the other party, acting as conditional cooperators with good governments, willing to pay their taxes, and as willing punishers with bad governments, engaging in tax evasion.

Lars Feld and Bruno Frey argue that the government has to proceed carefully in reacting to tax evaders to maintain conditional cooperation. They recommend leniency if the tax authorities uncover what is clearly just an unintended mistake, moderate punishment for small offenders, and harsh penalties for serious violations such as refusing to file a return. The first two maintain a spirit of conditional cooperation between the taxpayers and the government and the last captures the idea that people are willing punishers when their fellow citizens badly misbehave. Such considerations are completely absent from the Allingham/Sandmo model.[5]

FRAMING EFFECTS OR CONTEXT DEPENDENCE

Behavioral economists point out that people's choices are often affected by how various options are presented to them, that is, how they are framed. The public goods experiments reported in Chapter 6 provide an often-cited example. As we noted there, James Andreoni found that subjects are more likely to play the public good if they are told that playing the public good benefitted the other subjects (a positive frame) than if they are told that not playing the public good hurt the other subjects (a negative frame), even though the game being played is identical in both cases Andreoni (1995).

Framing effects abound in real-life situations as well. Indeed, one of the greatest triumphs of the behavioral economists was changing the framing of the 401K retirement plans. There is a grave concern in the United States

4. The ultimatum game is described in Rabin, op. cit., p.667. For a good introduction to the work of Fehr et al., on reciprocal altruism, see Fehr and Gachter (2000a) and Fehr and Gachter (2000b). A more detailed account can be found in Fehr and Schmidt.

5. Feld and Frey (2002). A good overview of tax compliance from a behavioral perspective is contained in McCaffery and Slemrod (2006), pp. 15–18. Their chapter also provides an excellent overview of the implications of framing effects and time inconsistencies for public sector analysis.

that a large percentage of people do not save nearly enough of their earnings over their working lives to support their retirements. Alicia Munnell and Annika Sunden reported that the median value of the financial assets of people with 55–64 years of age, those near retirement, was only $30,000 in 2004 Munnell and Sunden (2006). For this reason, the federal government established tax-deferred savings plans for employees, called 401(k) plans, as an incentive to save more for their retirement. Employers have the option of choosing one of two default options for participation under the 401(k) plans: automatic enrollment, in which employees are enrolled in the plan unless they specifically opt out, and optional enrollment, in which employees have to specifically choose to participate in the plan—the default option is nonparticipation. Offering automatic enrollment plans has an enormous positive effect on participation rates, an effect first uncovered by Brigitte Madrian and Dennis Shea in 2001. They studied the participation rates of employees of a large corporation that changed its 401(k) plan from optional to automatic enrollment effective on April 1, 1998. They found that the participation rates for a comparable cohort of the company's employees increased from 37% before the change to 86% after the change.[6]

William McCafferty and Joel Slemrod, in a review of the implications of behavioral economics for public sector issues, mention four framing effects that are potentially highly relevant to the design of tax and transfer policies: (1) People are more willing to accept a given increase in taxes if it is expressed as a percentage increase than as a dollar amount; (2) The so-called Schelling effect, in which people prefer both progressive transfer schedules and progressive taxes, even though a progressive transfer schedule is in effect a regressive tax within the income range over which the transfers are paid; (3) People prefer bonuses to penalties, such as a tax credit for replacing older appliances with new, more energy-efficient appliances rather than a penalty for retaining the older appliances; and (4) People prefer paying governments a penalty or fee than a tax of equal amount.[7]

The existence of framing effects is the most radical and potentially the most damaging challenge to the standard economic model because it implies that preferences are context dependent and therefore can be fundamentally inconsistent. Inconsistent preferences are difficult for any model to capture. This is particularly true for framing effects since economists and psychologists have such little understanding of why they occur.

These four anomalies are by no means a complete list of anomalies uncovered by behavioral economies. But they are sufficient to underscore how fundamentally behavioral economics is challenging the standard theory of consumer behavior. Douglas Bernheim and Antonio Rangel, two mainstream theorists who are trying to achieve a reconciliation of the standard theory with the insights of behavioral economics, list the following four items as the fundamental assumptions of the standard theory of consumer behavior. Following each item, we place in parentheses the anomalies we have discussed that violate the standard assumptions.[8]

> A1: Coherent preferences: Each individual has coherent, well-behaved preferences (framing effects).
> A2: Preference domain: The domain of each individual's preference rankings is the set of his/her lifetime, state contingent consumption paths, with future consumption discounted to present value by means of exponential discounting (prospect theory; present-biased preferences—quasi-hyperbolic discounting; social preferences).
> A3: Fixed preferences: Each individual's ranking of lifetime state-contingent consumption paths remains constant across time and states of nature. People's tastes are allowed to vary over time and across states, but in any given time or state they cannot question the decision made in another time or state (framing effects; present-biased preferences/self-control problems).
> A4: No mistakes: Each individual always selects the most preferred alternative from the feasible set (present-biased preferences/self-control problems; framing effects).

In summary, no element of the standard theory of the consumer has escaped attack from behavioral economics.

MAINSTREAM REACTIONS

As one might imagine, the behavioral challenges have met resistance from mainstream economists, for a number of reasons. Perhaps the most fundamental is the simplicity and usefulness of the standard model. George Stigler noted that the three main requirements of a model are accuracy of prediction, tractability, and generality. The standard model certainly meets the last two requirements and does reasonably well on the first score Stigler (1950).

6. Madrian and Shea (2001). The percentages cited are in Table IV, p. 1160.

7. McCaffery and Slemrod, op. cit., pp. 7–9. The last effect was especially ironic in the passage of the Affordable Care Act ("Obamacare") and its subsequent acceptance by the U.S. Supreme Court. The Obama administration specifically chose to refer to the payment that the uninsured would be assessed if they did not have insurance as a penalty, fearing that calling the payment a tax would doom its chances of being passed by Congress. Yet Chief Justice Roberts upheld the constitutionality of the payment by saying that it was essentially a tax and the government had the right to levy a tax on the uninsured.

8. Bernheim and Rangel (2005), pp. 5–11. The final section of this chapter discusses their approach for incorporating the behavioral anomalies into the standard theory.

Mainstream economists are generally sympathetic with Gary Becker's view that the standard assumptions of stable preferences and maximizing behavior, in combination with market equilibrium, have been useful in understanding all human behavior Becker (1976). The generality of the standard model also makes it possible to transfer insights gained in analyzing one situation to other situations that on the surface may seem to be quite dissimilar.

Useful is not the same as unerring, of course, and mainstream economists generally concede that the anomalies uncovered by the behavioral economists are systematic and important enough to be taken quite seriously. Nonetheless, they are reluctant to abandon the standard model unless the behavioral economists can develop an alternative general model of consumer behavior based on psychological foundations, and that has not yet happened. Khaneman and Tversky's prospect theory is the closest example of what the mainstream economists would be looking for and it falls far short of being a truly general model of behavior. Instead, behavioral economics appears to mainstream economists as a somewhat idiosyncratic one-by-one attack on the assumptions of the standard model without as yet any unifying foundations. One problem the behavioral economists face is that psychologists and psychiatrists have not developed models of much generality that explain under what conditions certain psychological problems such as ignoring relevant information are likely to occur. In short, it seems premature for mainstream economists to abandon the standard model.

A related point favoring the standard model in the context of a market economy is that agents who follow the standard model will necessarily outperform agents who exhibit anomalous behavior and eventually drive them from the marketplace. This is certainly true if markets are reasonably competitive. Therefore, the anomalies may ultimately not matter very much. This point brings little comfort to public sector economists, however, because there are no obvious corresponding mechanisms within the government sector to arbitrage the anomalies away.

A second reason for resistance is that the behavioral quest to find deep psychological foundations to the preferences that guide individuals' economic decisions is antithetical to the mainstream approach. The mainstream perspective focuses exclusively on the choices that people make which, through Samuelson's principle of revealed preference, reveal the underlying preferences behind those choices. Estimates of demand and factor supply curves based on an individual's choices can be integrated to reveal an indirect utility function lying behind those choices. Recall from Chapter 4, Dale Jorgenson's method of estimating a complete system of demand equations and then using the estimates to recover the transcendental log indirect utility function consistent with the estimated demand functions. The estimated utility function can then be used to make out-of-sample predictions of the individual's behavior, such as when a new tax changes some of the prices that the individual faces or, as in Jorgensen, as arguments in a flexible form of social welfare function to track changes in social welfare over time. In other words, the individual is assumed to behave *as if* he is maximizing the utility function derived from his choices, subject to a budget constraint, when making decisions. There is no need to assume that the individual actually carries out the maximization with that particular utility function in mind. Hence there is no need to understand what lies behind those preferences in any deep sense, whether psychological or otherwise. The revealed utility function is just a convenient and systematic way of summarizing how an individual arrives at the choices he makes.

A closely related mainstream concern is that if choices do not reveal preferences, then what does? If one assumes that the utility functions revealed by individuals' choices are not their true preferences, then one has to add parameters to, or otherwise modify, the utility functions based on appeals to nonchoice data. Examples would be a presumed underlying psychological motivation that can be parameterized, self-reported data on preferences, and neurological brain images obtained in various mental states—distracted vs. focused, distressed vs. calm—when making decisions. Mainstream economists are generally skeptical about the reliability of nonchoice data as a window to true preferences and they worry that such data might not be able to be measured accurately. One reason that quasi-hyperbolic discounting has gained some traction in the mainstream is that the key parameter β can be estimated. Still another concern is adding more parameters to the estimation of preferences when the estimation of utility functions under the standard model already requires a number of assumptions about structure and parameterization.

The issues actually run deeper than finding relevant, reliable, and measurable nonchoice data. Suppose one accepts the behavioral presumption that economic decisions or choices are often made based on preferences that differ from true preferences. As the behavioral economists put it, individuals have *decision preferences* that determine their choices and separate *experience preferences* that are their true underlying preferences. The usual approach in this case is to modify one part of the standard model to capture the presumed anomaly, estimate the model such that the modification best explains the observed choices, and then let the unmodified model represent the true preferences. For example, assuming that individuals have present-biased preferences in making decisions, estimate the β under the assumption of quasi-hyperbolic discounting such that it best fits the self-control problem being studied, and then assume that $\beta = 1$ (exponential discounting) to represent the true preferences. This approach raises a natural question, however: If individuals get any one part of the

standard maximizing model wrong, why should they follow the rest of it? Mainstream economists ask behavioral economists to develop general models that incorporate a number of important behavioral anomalies all together, rather than proceeding one anomaly at a time. That goal seems to be a long way off, however.

Still another mainstream reservation is that a particular anomaly can often have a number of possible psychological motivations. The success of changing the default option on 401K retirement plans is a case in point. Under the original default option of not joining, did people choose not to join because: They had present-biased preferences for current consumption? They were simply inattentive? They tend to procrastinate? They did not understand the importance of saving early on for retirement? One can imagine how changing the default option might have changed their behavior under any one of these explanations.

Finally, many of the behavioral anomalies were discovered in experimental settings. Mainstream economists wonder if they will hold up as the research turns more to behavior in actual situations. We will pursue this point in more detail when discussing Khaneman and Tversky's prospect theory.

POSITIVE AND NORMATIVE PUBLIC SECTOR ECONOMICS

The behavioral anomalies have a direct effect on both the positive and normative analysis of public policies. Their effects on positive analysis are easier for mainstream economists to accept. It is certainly reasonable to assume that a better understanding of the psychology underlying individuals' choices would lead to a better understanding of those choices. This in turn would help governments to design policies that are more likely to meet the goals they are trying to achieve. For example, the U.S. federal government has long tried to encourage people to save more for their retirement. One way to achieve this is to replace the personal income tax with a consumption tax. Another way, and the one chosen so far, is to offer people tax-free retirement saving devices such as Individual Retirement Accounts (IRAs) and 401Ks. If people save too little for retirement because of anomalies such as inattention to long-run planning or present-biased preferences, then the latter may be the more effective policy. Advertising the retirement saving options calls attention to the importance of saving and their tax-free status may help to overcome the present bias to consume now instead of saving for a distant future. Moreover once these saving instruments were offered, changing the default option to automatic enrollment overcame a framing anomaly, rather dramatically.

In general, the behavioral anomalies have opened up a whole new line of public sector analysis. Public sector economics has traditionally been concerned with correcting market failures. This is achieved in three ways: (1) Having the government offer desired goods and services that would otherwise not be provided by the market economy, e.g., nonexclusive goods, hard-case decreasing cost services, universal medical insurance; (2) Altering consumers' budget constraints and firms' profit functions by means of taxes, subsidies, and transfers, e.g., income and sales taxes to finance publically provided services, Pigovian taxes to correct for externalities, providing transfer payments to the poor; and (3) Providing information, e.g., subsidizing basic scientific research, and testing to ensure that food and medicines are safe and monitoring the safety of the workplace. Behavioral economics opens up the possibility of designing policies to correct failures of individual decision-making, that is, to correct the behavioral anomalies, a line of inquiry commonly referred to as behavioral public finance. The methods employed are such as noted above with the IRA and 401K saving options: countering biases and inattention and altering the frames under which people make decisions. Richard Thaler and Cass Sunstein called these decision-altering devices "nudges," a term that has stuck in the public sector literature Thaler and Sunstein (2008). Mainstream economists generally support this line of research.

At the same time, psychological motives that modify the standard assumptions have direct consequences for normative analysis, and our impression is that mainstream economists are uncertain about how to view the normative implications. The normative policy prescriptions for public expenditures and taxation developed throughout this textbook rely on the standard assumptions along with, in almost all cases, perfectly competitive markets. Change the underlying assumptions of consumer theory to accommodate one or more of the behavioral anomalies, and the normative policy prescriptions change as well. For example, transforming the general equilibrium quantity model used to analyze first-best public expenditure and tax theory in Parts I and II of the text into a general equilibrium price model used in second-best tax and expenditure theory in Part III of the text assumed the standard assumptions and perfect competition. The general equilibrium price model would change under modified assumptions and along with it all the second-best policy prescriptions change. This is not a problem in principle. Quantity models can be turned into price models under assumptions about consumer behavior that differ from the standard assumptions. But without a general behavioral theory of the consumer to replace the standard theory, the proper way to make the transformation from quantity models to price models is not at all clear.

A related issue concerns the specification of social welfare. Presumably the objective of public sector analysis from the behavioral perspective remains the maximization of a Bergson–Samuelson social welfare function whose

arguments are the individuals' utility functions. But what individual utility functions should a social planner use: those that incorporate parameters that capture various behavioral anomalies or those that embody only the standard assumptions? The answer likely depends on whether particular behavioral anomalies are viewed as mistakes in the sense that they are clearly counter to rational individuals' self-interest or they are embedded in individuals' underlying psychological makeup and should not be viewed as mistakes. If it is the latter, then one could argue that decision utility and experience utility are one and the same; the standard theory is simply wrong. The utility based on apparently anomalous choices from the mainstream perspective is the appropriate utility for the social planner to use. If the anomalous behavior is viewed as a mistake, however, then it is not clear how to proceed.

Consider present-biased individuals with self-control problems who appear to employ quasi-hyperbolic discounting in making choices over time. Some economists argue that the social planner should assume correct discounting with $\beta = 1$ for purposes of calculating social welfare, but assume that $\beta < 1$ when designing policies to offset the present bias. This position assumes that the present bias ($\beta < 1$) is simply a mistake. Other economists wonder why actual decisions about the consumption path with $\beta < 1$ should not count for something in terms of social welfare since the present-biased consumption path clearly brings some utility to the individual. The question, though, is how much it should count. Still other economists see multiple selves in each individual, because the present bias arises on a rolling basis for each time period as people decide between present and future allocations. This leads to the problem of how the social planner should aggregate the utilities of each person's multiple selves, the standard aggregation problem in social welfare analysis. In short, how to define the social welfare maximum in policy analysis in the presence of behavioral anomalies remains an open question among mainstream economists.

The uncertainty about what the social welfare function should be is just one manifestation of a more general and fundamental problem. The idea that people may have inconsistent preferences or make mistakes that are clearly against their own self-interest raises a truly fundamental issue for the mainstream normative theory of the public sector. It undermines the principle of consumer (producer) sovereignty that people are the best judges of their own self-interest, which is the foundational value judgment on which the theory rests. Consumer sovereignty is the basis for defining the efficiency norm as pareto optimality and representing the equity norm of distributive justice in terms of the Bergson–Samuelson individualist social welfare function. Consumer sovereignty is also the basis of the government-as-agent principle, that the preferences of the individuals, not those of the public officials, are what matter in determining what functions government should perform and how they should proceed within each function. Take away consumer sovereignty, and the entire mainstream normative public sector theory falls apart.

There is a practical problem as well. The notion that individuals' choices can be overridden when determining social welfare leads to the slippery slope of policy makers deciding what is the "natural" thing for people to do or believe in certain situations. What standard of proof is necessary to accept some behavior as "natural" if it is other than what people choose to do? And what is the normative significance of a subset of people in their role of public officials determining how others should have behaved? Also, if people are susceptible to framing, then how public officials present their policy proposals may carry more influence than the underlying economic merits of the policies. The normative uncertainties are compounded if the public officials themselves are afflicted with irrational anomalies when making decisions, in which case the biased are leading the biased on a march toward, well, toward what?

Small wonder, therefore, that mainstream economists appear unwilling to abandon the standard first- and second-best normative policy prescriptions that appear throughout this textbook despite conceding the existence of systematic and important behavioral anomalies. And they are unlikely to do so until and unless a convincing general theory of behavioral economics is developed to replace the standard theory.[9]

PROSPECT THEORY

KT's prospect theory deserves special attention because it is the most fully developed of the behavioral theories. They have conducted a large number of experiments in order to explain how people behave under uncertainty. The experiments convinced them that the standard model of expected utility maximization is completely wrong and they proposed an alternative theory, which they called prospect theory.[10]

According to the standard theory, individuals have concave utility functions defined over their wealth, W, with

9. There are a number of excellent discussions of behavioral economics by mainstream economists, in which the points made in this section are discussed in more detail. A good beginning set of readings would be: Fudenberg (2006); E. McCaffery and J. Slemrod, *op. cit.*; B. D. Bernheim and A. Rangel, op. cit; Postlewaite (2011); and Pesendorfer (2006). E. McCaffery and J. Slemrod are a particularly good source on the implications of behavioral anomalies for mainstream public sector theory.

10. Kahneman and Tversky discovered some anomalies with their original theory, which are corrected by postulating a cumulative version of their theory, in which the gains and losses are interpreted as gaining at least X or losing at least Y. The updated version appears in Kahneman and Tversky (1992). The analysis in this section, and in Barbaris' review of prospect theory, is based on their cumulative prospect theory.

the concavity of their utility functions reflecting risk aversion. Suppose that an individual has initial wealth W_0 and faces an uncertain situation with N possible draws or states of nature. He makes a decision that results in a gain or loss of X_i in state of nature i with objective probability p_i, $i = 1, \ldots, N$, which the individual knows. According to the standard theory, the individual maximizes his expected utility over the levels of wealth he would have in each state of nature:

$$\text{Max. } E(U) = \sum_{i=1}^{N} p_i U(W_0 + X_i).$$

The individual updates the p_i using Bayes' rule as new information on the probabilities becomes available.

KT found that their subjects do not do this at all. Instead, they behave as if they are maximizing a function

$$\text{Max} \sum_{i=1}^{N} w_i V(X_i).$$

They call V, a value function and the w_i, a set of decision weights attached to each state of nature rather than the objective probability of that state occurring. V and the w_i have the following four properties (the terms are those of KT).

1. *Reference dependence*—Notice, first, that V is a function of the X_i, the changes in wealth but not the levels of wealth. Moreover the changes are defined relative to a reference point, hence the term reference dependence. The reference point might not be the initial level of wealth W_0.
2. *Loss aversion*—Individuals are much more sensitive to losses from the reference point than to gains of an equal amount. The value function is given by Fig. 25.1, which assumes for simplicity that the initial wealth is the reference point.

The aversion to loss is very strong indeed. V is steeper at each amount of loss relative to the corresponding amount of gain. Moreover, V has the usual concavity over the entire range of gains but is convex over at least moderate amounts of loss. Individuals are risk lovers over this range. KT found, for example, that subjects typically reject the following gamble: a gain of \$110 with $p = 1/2$ or a loss of \$100 with $p = 1/2$. They place a higher negative value on the loss of \$100 than the positive value of a gain of \$110. In contrast, the standard theory of utility maximization predicts that most subjects would accept the gamble because the gains are so small relative to their initial wealth that they should be essentially risk neutral with respect to the gamble. In addition, experimental subjects typically favor a gamble of losing \$1000 with $p = 1/2$ to a loss of \$500 with certainty. The certain loss is so painful that they prefer to take the risk. In the standard theory, risk-averse individuals would favor taking the \$500 loss over the gamble with an expected loss of \$500. In prospect theory, people are consistently risk averse only with respect to gains.

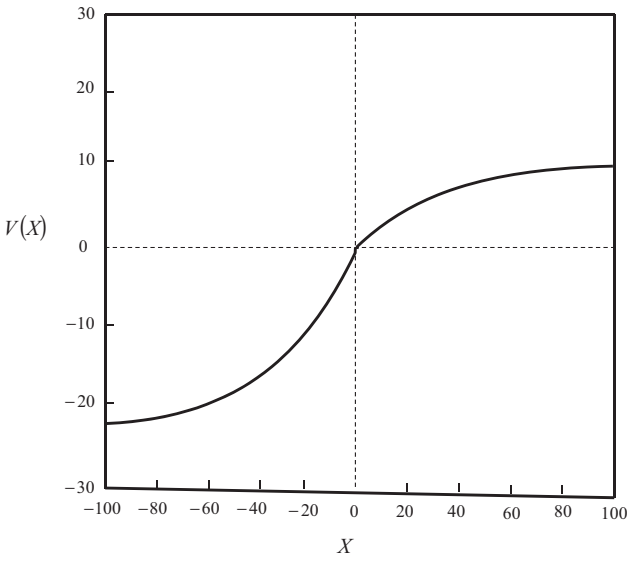

FIGURE 25.1

The rejection of the 50/50 (−\$100, +\$110) gamble indicates that people also compartmentalize gains and losses rather than considering the gains and losses in the context of their overall wealth or future investment opportunities. Shlomo Benartzi and Richard Thaler referred to this as keeping separate mental accounts for things that are actually entirely interdependent, a kind of narrow framing Benartzi and Thaler (1995). A common example is refusing to sell a house when sellers receive lower offers than they had hoped to receive (the reference point) even though they have excellent investment opportunities elsewhere that they forego by holding on to the house.

3. *Diminished sensitivity*—V becomes flatter in both directions as the X_i increase, suggesting that individuals have diminished sensitivity to additional gains or losses once they have experienced gains and losses. For instance, replacing a \$100 loss with a \$200 loss entails more utility loss than replacing a \$1000 loss with a \$1100 loss.
4. *Decision weights*—KT found that the subjects apply decision weights to the various states of nature that are related to the objective probabilities of each state occurring, but that they systematically overweight the tails of the distribution. The pattern is illustrated in Fig. 25.2, in which the objective probabilities are on the horizontal axis and the decision weights are on the vertical axis.

The two are the same under certainty, with the p_i equal to zero or one. Otherwise the decision weights exceed the objective probabilities below $p_i = {}^1/_2$ and are less than the objective probabilities above $p_i = {}^1/_2$ (they are equal

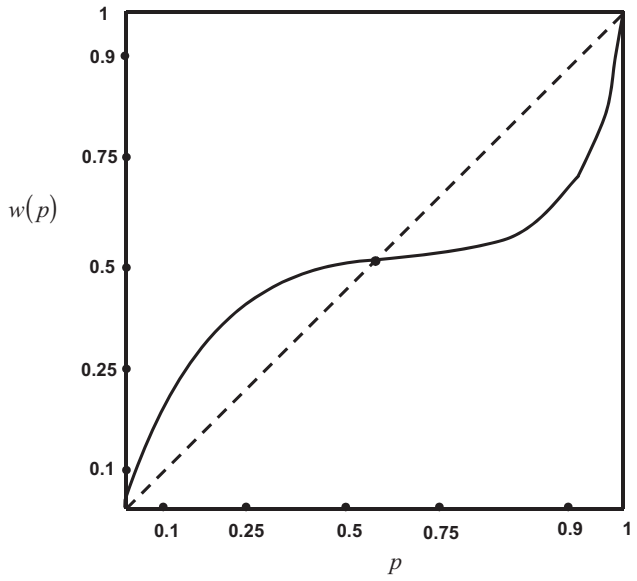

FIGURE 25.2

at $p_i = {}^1/_2$). Moreover, the divergence between the two increases until the objective probabilities are very close to zero and one. That is, people are especially prone to overweighting extreme outcomes relative to their objective probabilities, at least until the extremes are essentially certain. KT suggest a weighting function of $w(p) = \frac{p^\delta}{(p^\delta+(1-p^\delta))^{\frac{1}{\delta}}}$, with $\delta = 0.65$, $w(p) = p$ if $\delta = 1$, the dotted line in Fig. 25.2.

KT observe that the overweighting of extreme outcomes can explain why people simultaneously gamble and take out insurance. Given the weighting function, people prefer a chance of winning \$5000 with a probability of $p = 0.001$ to a gain of \$5, but prefer a certain loss of \$5 to a probability of $p = 0.001$ of losing \$5000. These preferences could not occur simultaneously under the standard theory of expected utility maximization, regardless of whether individuals are risk averse, risk neutral, or risk lovers.

KT also note that the decisions weights do not represent mistakes in the sense that their subjects do not understand the objective probabilities. They fully understand the objective probabilities of various outcomes in the experiments. They simply choose not to use them in the standard way in making their decisions.

APPLYING PROSPECT THEORY

In his review of prospect theory, Nicholas Barberis notes that other researchers using ever more sophisticated experiments have confirmed the four elements of prospect theory. As noted earlier in the chapter, Barberis believes that prospect theory best explains how subjects behave in experimental settings involving risk.[11] That said, the theory has not as yet been widely applied to actual settings, although research along those lines is beginning to gain some momentum. Most applications to date have been in finance and insurance, which makes sense given that it is a theory of behavior under uncertainty. It has had very little impact on public finance.

The main difficulty in applying the theory has been in capturing the property of reference dependence in real-life settings. Reference points are easy to define and manipulate in experimental settings, but often very difficult to determine with much confidence in actual situations. The difficulties are such that the vast majority of empirical research still favors expected utility maximization as the underlying theoretical foundation. This may change, however, should more convincing ways of defining actual reference points evolve, given prospect theory's overwhelming success in laboratory settings.

The one potential application we will consider here concerns annuities. As noted in Chapter 21 on social insurance, private annuity markets are exceedingly thin in the United States, such that annuities are too expensive. Peter Diamond uses this fact in support to his position that the public annuities under the Social Security System should be retained. Diamond attributes the thinness of the annuity markets to a simple mistake: People do not understand that annuities are the cheapest way to provide a given stream of income during the retirement years Diamond (2004). Prospect theory offers two other possibilities for avoiding private annuities: loss aversion and overweighting of extreme outcomes. Purchasing a retirement annuity is a gamble. Assume that the purchase price is actuarially fair, equal to the present value of the annuity income stream to the expected year of death. If the purchaser dies before reaching the average life expectancy, he or she loses. If the purchaser lives longer than the average life expectancy, he or she gains. Under loss aversion, the possible loss of a shorter-than-expected life outweighs the possible gain of a longer-than-expected life. The overweighting of extreme outcomes enhances the hesitancy to purchase an annuity because of loss aversion. Suppose that a quite healthy person at 65 years of age is considering buying an annuity. Conditional on reaching age 65, the life expectancy in the United States is in the mid-80s. The probability of the person dying shortly after buying the annuity is quite small. But the person overweights the probability and, combined with loss aversion, foregoes the annuity.

11. Our presentation of prospect theory closely follows that in N. Barberis, op. cit., pp. 2–9. Figs 25.1 and 25.2 reproduce Figs 1 and 2 on pp. 33 and 34, and the KT estimate of δ in the weighting function is on p. 34. Barbaris also considers applications of prospect theory in the literature in the areas of finance, insurance, consumption-savings decisions, labor supply, and industrial organization, among others.

Why, then, are people reluctant to purchase retirement annuities? Is their reluctance a mistake, *a la* Diamond, that could possibly be overcome with an educational campaign about the advantages of retirement annuities? Or is it a more fundamental reluctance caused by people's underlying psychological tendencies, and thus much more difficult to overcome? The answer is unclear, but it likely matters in the debate about whether public pension systems should be privatized.

NUDGES AND STANDARD POLICY PRESCRIPTIONS

Sendhil Mullainathan, Joshua Schwartzstein, and William Congdon (MSC) developed a model to explore the effects of behavioral anomalies on public sector issues. Their model is particularly useful for seeing how nudges designed to reduce or remove the anomalies interact with standard public policy prescriptions.[12]

The model consists of a continuum of atomistic individuals with identical preferences and the same fixed income Y. Each individual decides whether to take a discrete action a, $a = [0,1]$. The cost of the action is its price, p, which reflects the marginal cost of the action. The marginal cost is constant no matter how many people take the action. The benefits of the action, b, are distributed across the individuals according to the cumulative density function, $F(b)$.

Standard agents with no behavioral anomalies take the action if $b > p$. Therefore, the number of agents who take the action is $A^S(p) = 1 - F(p)$. The so-called behavioral agents make an error, e, in assessing the benefits of the project, such that they take the action if $b + e > p$. MSC view the error as just that, a mistake that lowers the behavioral agents' utility. Therefore, utility that includes the standard net benefit of the action, $b - p$, is the true or experience utility and the utility that includes the behavioral net benefit of the action, $b + e - p$, is the decision utility. The number of behavioral agents who take the action is $A^B(p) = 1 - F(p - e)$. $A^B(p) < A^S(p)$, if behavioral agents underestimate the benefit ($e < 0$) and vice versa, if they overestimate the benefit.

The government levies a per-unit commodity tax t on the action. Given the assumption of constant marginal cost, the price of the activity rises by the full amount of the tax. The tax revenue collected is $R(t) = tA(t)$. (We drop the superscript B or S to indicate that the revenue equation applies to either type of agent, providing all agents are the same type.) The government uses the tax revenue to provide a per-unit subsidy, T, to each person and possibly to finance some other goods. These other goods are unspecified and have no effect on the individuals' utilities. The government is subject to an overall budget constraint $G(t,T) = 0$. Thus, in general, $T = g(R(t))$, $0 < g' \leq 1$.

12. Mullainathan et al. (2012). The *Annual Review of Economics* is online at economics.annualreveiws.org. This section follows closely the presentation of the model in their paper.

STANDARD AGENTS ONLY

Consider, first, the model with only standard agents. The utility of an individual who does not take the action is $U = U(Y + T - \mathrm{l}A^S)$. The term $\mathrm{l}A^S$ captures the possibility that taking the action generates an externality l on each individual. l is fixed, independent of the number of people, A^S, who take the action. The utility of an individual who takes the action is $U = U(Y + T + (b - p) - \mathrm{l}A^S)$. Incorporating both kinds of individuals by means of the parameter $a = [0,1]$, $U = U(Y + T + a(b - p) - \mathrm{l}A^S)$.

Since all individuals have identical preferences, a natural goal for the government is to maximize expected utility. Its single policy tool is the tax rate, t, applied to the action. Therefore, social welfare is

$$\begin{aligned} W^S(t) &= E^S(U(Y + T(t) + a(b - p(t)) - lA^S(p(t))) \\ &= \int_{-\infty}^{p(t)} U\big(Y + T(t) - lA^S(p(t))\big)dF(b) \\ &\quad + \int_{p(t)}^{\infty} U\big(Y + T(t) + (b - p(t)) \\ &\quad - lA^S(p(t))\big)dF(b). \end{aligned}$$

But $T(t) = g(R(t)) = g(tA^S(t))$. Therefore,

$$\begin{aligned} W^S(t) &= \int_{-\infty}^{p(t)} U\big(Y + g\big(tA^S(t)\big) - lA^S\big(p(t)\big)\big)dF(b) \\ &\quad + \int_{p(t)}^{\infty} U\big(Y + g\big(tA^S(t)\big) + \big(b - p(t)\big) \\ &\quad - lA^S\big(p(t)\big)\big)dF\big(b\big). \end{aligned}$$

Maximizing social welfare with respect to t,

$$\begin{aligned} \frac{dW^S}{dt} &= -E^S[U'(C)]\big[lA^{S'}\big(t\big) - g'\big(R\big(t\big)\big)\big(A^S\big(t\big) + tA^{S'}\big)\big] \\ &\quad - E^S[U'(C)]_{|a=1}A^S(t)\frac{dp}{dt} \end{aligned} \tag{25.2}$$

The derivative incorporates the envelope theorem. $b = p(t)$ for the marginal individual. Therefore, by the envelope theorem, the unit changes in the upper and lower limits of the two integrals have no effect on the utility of the marginal individual.

Dividing Eq. (25.2) by $\frac{dW^s}{dY} = E^s U'(C)$ to express the derivative in terms of units of income rather than utility,

and recalling that p rises by the full amount of the tax, so that $dp/dt = 1$,

$$\frac{dW^S}{dt}\Big/\frac{dW^S}{dY} = [g'(R(t))(A^S(t) + tA^{S'}) - lA^{S'}] - A^S(t)_{|a=1} \tag{25.3}$$

Adding and subtracting $tA^{s'}$ and rearranging terms yields

$$\frac{dW^S}{dt}\Big/\frac{dW^S}{dY} = [t + (g'(R(t)) - 1)t - l]A^{S'} + (g'(R(t)) - 1)A^S(t)_{|a=1} \tag{25.4}$$

MSC write the RHS of Eqn. (25.4) as $(t + ME(t))A^{S'} + TV^S(t)A^S$, where

$ME(t) = (g'(R(t)) - 1)t - l$ and $TV^S(t) = g'(R(t)) - 1$. $ME(t)$ is the marginal external effect of taking an action, consisting of the marginal effect on the government budget constraint, that not all the extra tax revenue goes to the transfer payment, and the marginal effect of the externality. $tA^{S'}$ is the usual marginal loss from increasing a tax, equal to the tax rate times the change in the quantity of the taxed good. Together, therefore, $(t + ME(t))A^{S'}$ represents the inefficiencies from increasing price above marginal cost, the excess of the marginal social cost over the marginal social benefit of taking the action. TV^s represents a redistributional motive of transferring income from those who take the action to the entire population. Examples could be transferring money from those with low social marginal utility to those with high social marginal utility, or providing social insurance, or raising revenue to finance public goods with high social value.

BEHAVIORAL AGENTS ONLY

This framework is useful for exploring the effects of some but not all behavioral anomalies. It cannot be used to analyze loss aversion because utility is specified in terms of levels of income and not changes relative to a reference point. Social preferences, representing a concern for how other people behave, are also ruled out. So too are errors caused by forgetfulness, since forgetfulness is a random event and the errors in the MSC model occur with certainty. But it can capture a number of common and important anomalies, such as present bias/self-control issues, inattention to components of price that are not obvious, a common problem with taxes, and false beliefs or overconfidence regarding beliefs.

The model with behavioral agents sets up exactly as the standard model, with two important differences. First A^B, the number of behavioral agents who take the action, replaces A^S, and social welfare is denoted as W^B. Second, the marginal behavioral agent who is just indifferent to taking the action, equates $b - p(t)$ to $-e$, not to zero. Therefore, the envelope theorem does not apply to that agent in the derivation of $\frac{dW^B}{dt}$. As the upper and lower limits of the two expected utility integrals change by one unit, the marginal agent undergoes a change of utility equal to $U(Y + T - \mathrm{l}A^S) - U(Y + T + e - \mathrm{l}A^S)$. MSC convert this error term to income units by dividing by $\frac{\partial W^B}{\partial Y} = E^B U'(C)$,

$$\overline{e} = \frac{U(Y + T - lA^S) - U(Y + T - e - lA^S)}{E^B[U'(C)]} \tag{25.5}$$

MSC refer to $\overline{e}$ as the marginal internality, labeled MI(t), the damage the marginal agent does to him or herself. Notice that if utility is linear, then $\overline{e} = e$.

Other than A^B rather than A^s and the marginal internality MI(t), there is no difference between $\frac{dW^s}{dt}$ and $\frac{dW^B}{dt}$. Therefore,

$$\frac{dW^B}{dt}\Big/\frac{dW^B}{dY} = [t + ME(t) + MI(t)]A^{S'} + TV^B A^B \tag{25.6}$$

The behavioral agent model provides a convenient framework for analyzing policy nudges to reduce or offset internalities. Before doing so, however, consider one result that follows immediately from Eqn. (25.6). Suppose $\mathrm{l} = 0$—there are no externalities, and $g'(R(t)) = 1$—all marginal changes in tax revenue change transfers T by the same amount. Under these two conditions, $\mathrm{ME}(t) = 0$, as does $\mathrm{TV^B}$. Therefore, setting $t = -\mathrm{MI}(t)$ generates $\frac{dW^B}{dt} = 0$. Pricing the internality is optimal. This is the behavioral justification for sin taxes to offset the present bias or self-control problems associated with smoking and alcohol consumption.

Nudges

Suppose the error is differentiable in nudge n. Then the government has the option of affecting e directly through the nudge and possibly improving welfare. Common examples of nudges are changing the default option on 401Ks, convincing people to do something by demonstrating that this is what others do, providing better information about the effective tax rates people face under a given tax (making taxes more salient), and simplifying enrollment procedures for people who are eligible for certain transfer programs. Given nudge n, the error e reduces to e_n, and behavioral agents now take the action if $b + e_n > p$.

Given the existence of a tax t, the effect of a marginal change in the nudge n in the behavioral model is

$$\frac{dW^B}{dn}\Big/\frac{dW^B}{dY} = \frac{dA_n^B}{dn}\mathrm{MI}_n(t) + \frac{dA_n^B}{dn}\big(t + \mathrm{ME}(t)\big), \tag{25.7}$$

where $\mathrm{MI_n(t)} = -\overline{e}_n(t) = \frac{U(Y+T-\mathrm{l}A^B) - U(Y+T+e_n(b_n^B,t)-\mathrm{l}A^B)}{E^B[U'(C)]}$

Notice that the direct redistribution effect, TV^B is absent from Eqn (25.7) because the nudge has no effect on the price p. Nudges affect the government's budget only indirectly by changing the number of people who take the action, which changes tax revenues through the $(t + \text{ME}(t))$ term.

Equation (25.7) yields a number of immediate insights about the relationship between standard mainline public policies and nudges to remove or reduce behavioral errors. First, suppose there are no externalities and the government can achieve the first best in the standard case, such that $t = 0$. For example, all taxes and transfers are lump sum. Then the second term in Eqn (25.7), $(t + \text{ME}(t))$, is zero and the government can achieve the optimum, $\frac{dW^B}{dn}/\frac{dW^B}{dY} = 0$, by using nudges to eliminate the error.

The use of nudges is more problematic in a second-best environment, however. In the presence of distorting taxes, nudges that improve behavioral agents' utility may have the overall effect of lowering social welfare. This is illustrated by Eqn (25.7). The first term is necessarily positive because A^B and e move in the same direction. If individuals underestimate the benefit of the action ($e < 0$), then the nudge increases the value of the error (e becomes less negative) and more people take the action. Conversely, A^B and e both decrease in response to the nudge if people overestimate the benefit ($e > 0$). The second term, which captures the marginal external effect of the nudge, can be positive or negative, however. It will be positive if and only if the more biased agents are more likely to take an action that is socially harmful on the margin. The overall effect is an example of the Lipsey/Lancaster theorem of the second best. Nudges that eliminate errors and are optimal in a first-best environment may not be social welfare improving in a second-best environment.

MSC offer a number of potentially important examples for which nudges and standard second-best policy prescriptions may work at cross-purposes, including the following:

People may overestimate the probability of having their tax returns audited. Providing accurate information about the auditing probabilities may induce some people to evade their taxes. This may make most of the new evaders better off but society worse off because of the lost tax revenues.

As we saw in Chapter 20, private and government insurance policies often require co-payments to offset the effects of moral hazard under private information. Otherwise some people may overuse the insurance under full insurance. At the same time, some of the insured may underestimate the advantages of taking certain medicines or having certain routine preventative procedures done. A nudge in this instance may take the form of a campaign to teach them the benefits of the medicine or the procedures. If these people are at the margin when a co-payment is introduced or increased, then the co-payment and the campaign work at cross-purposes to one another.

In Chapter 19, we described the unambiguously negative work incentives in the second and third phases of the Earned Income Tax Credit, when the credit is constant (second phase) and decreasing (third phase). If some people do not realize they are eligible for the credit in these two phases, a campaign to make them aware of the credit may induce them to apply for the credit and simultaneously reduce their labor supply. They are undoubtedly better off, but society may be worse off because of the reduced supply of labor.

A MIXTURE OF STANDARD AND BEHAVIORAL AGENTS

MSC also analyze the realistic case when there are both standard and behavioral agents. An obvious goal in this case is to find incentives that reduce or eliminate the errors of the behavioral individuals but have no effect on the standard individuals. Rather than reproduce the full mixed agent model, we can illustrate an optimal policy with a very simple representation of self-control/present-biased individuals that MSC use to illustrate a number of points.

Suppose that individuals engage in an activity such as cigarette smoking that has benefits and costs spread over two periods. For simplicity, the discount rate for the second period is assumed to be zero. The benefits of the activity have two components, a benefit v that occurs in the first period and a delayed cost h that occurs in the second period. The benefit v is distributed according to the cumulative density function, $F(v)$. The delayed cost h is a constant. Think of v as the immediate pleasure from smoking cigarettes and h as future ill health caused by smoking. The cost or price of the activity, p, is experienced in the first period and is constant.

The standard individuals undertake the activity if $v - h > p$. The behavioral individuals with a self-control/present-bias problem engage in semihyperbolic discounting of the future with $\beta \in (0, 1)$. They undertake the activity if $v - \beta h > p$. Notice that the error $e = h(1 - \beta)$ in terms of the MSC formulation of the error term.[13]

The government can correct the behavioral error while leaving the standard individuals unaffected with a two-period tax policy, setting $t_1 = h$ and $t_2 = -h$. The standard individuals now undertake the activity if $v - h = p + t_1 + t_2 = p + h - h = p$. They are unaffected by the taxes. The behavioral individuals now undertake the activity if $v - \beta h > p + h + \beta(-h)$.

13. $v - h + e = v - \beta h$. Therefore, $e = h(1 - \beta)$.

Rearranging terms and canceling, we obtain $v - h > p$. The two-period tax policy removes the present-bias error and is welfare maximizing in an otherwise first-best environment.

CAN MAINSTREAM AND BEHAVIORAL ECONOMIC THEORY BE RECONCILED?

Mainstream theorists B. Douglas Bernheim and Angelo Rangel, have recently proposed a way of incorporating behavioral anomalies into mainstream theory that has gained the attention of both mainstream and behavioral economists. The appeal of their approach is hardly surprising because it incorporates both points of view. On the one hand, being mainstream theorists, they insist on maintaining the central premise of the mainstream theory that individuals' choices have to be the foundation for any inferences about their preferences and welfare. On the other hand, they recognize that the various behavioral anomalies can lead to inconsistent choices that mask individuals' true preferences. But they are confident that advances in psychology and neuroscience will help eliminate some of the inconsistent choices and give policy makers a better sense of what individuals truly prefer. This avoids the slippery slope of basing policy decisions on what government officials think people ought to prefer when their preferences are unclear. In other words, the Bernheim/Rangel (BR) approach is consistent with the idea that good psychology makes good economics, which is the central premise of behavioral economics.

Following BR, the standard approach to consumer behavior can be represented as follows. Define ℵ as the set of all elements that an individual is interested in choosing and thus the set of elements over which the individual's preferences are defined. The elements could be bundles of goods and services, lotteries over these bundles, a sequence of bundles over the individual's lifetime, anything that could be an object of choice. Within ℵ is a subset X of the elements that the individual is constrained to choose from in a given situation, which BR refer to as the standard constraint situation (SCS). The constraint set X depends on the information available to the individual and possibly a given amount of resources if some of the individual's resources are fixed. A given constraint set could include all the elements of ℵ but it may not. The choices made under various SCSs are then used to estimate the individual's preferences over ℵ by using the revealed preference relations of preference and indifference. The revealed preference relations represent a complete ordering over all the elements in ℵ under standard assumptions. Therefore, the individual's preferences and associated utility or welfare are just a summary of the choices that the individual has made, a summary based on the revealed preference relations.

BR propose that the standard framework be maintained in the presence of behavioral anomalies, with the following modification. Define a generalized constraint situation (GCS), G, consisting of a constraint set X and an ancillary condition, d, such that $G = (X,d)$. An ancillary condition describes the environment of context in which a choice is made, and is a potential source of a behavioral anomaly. It could be a default option, a particular way of framing a choice, a time period in which a decision is made (relevant for present-biased individuals), an individual's state of mind (calm, nervous) when the choice is made, a whole range of possible conditions that might be present in the given situation. Observe the choices made under the GCSs and then apply the revealed preference relations to determine the underlying preferences that the choices represent.

Unfortunately, the revealed preference relations applied to the choices made in the presence of behavioral anomalies might not generate a complete ordering of preferences. To see the kinds of problems that can arise, consider the four revealed preference relations that can apply over the GCSs.

Let x and y be two elements within the GCSs.

xRy means that x is no worse than y. It says that if x and y are both available in a GCS, x is sometimes chosen and y is never chosen unless x is as well.

xIy means that x is indifferent to y. It implies xRy and yRx. If x and y are both available in a GCS, then either both are chosen or neither is chosen.

xPy means that x is weakly preferred to y. It says that if x and y are both available in a GCS, x is sometimes chosen and not y. Otherwise, either both are chosen or neither is chosen.

xP*y means that x is strictly preferred to y. It says that if x and y are both available in a GCS, x is sometimes chosen and not y. Otherwise, neither is chosen.

A result of particular importance in welfare analysis is the identification of strict individual welfare optima, choices that cannot be weakly improved upon. x is a strict individual welfare optimum in X if for each $y \in X$ other than x, one of two conditions hold: either x is chosen and y is not for some (X,d) or there is no (X,d) for which y is chosen but x is not with x present. Under either condition, it cannot be that for some $y \notin X$, yPx. An immediate corollary is that x is a strict individual welfare optimum if x is the unique choice for some (X,d).

Suppose two GCSs are $\{(x,y), \mathrm{d}'\}$ and $\{(x,y), \mathrm{d}''\}$. If d' and d'' represent two different frames, then it is quite possible for the frames to generate a choice reversal, such that x is chosen under $\{(x,y), \mathrm{d}'\}$ and y is chosen under $\{(x,y), \mathrm{d}''\}$. In this case the ordering is not complete.

Nor will the ordering necessarily be transitive. BR offer the following example with inconsistent preferences over three elements (but without reference to ancillary conditions). The choice is given following each set:

(x, x_2); x_1 (x_2,x_3); x_2 (x_1, x_3); x_3 (x_1, x_2, x_3); x_1, x_2, x_3

Under the revealed preference relations, $x_1 P x_2 P x_3 P x_1$, the ordering is intransitive. BR are able to show, however, that P^* is acyclic under the assumption that all subsets of the choice elements are considered. This permits identification of most preferred alternatives and hence a welfare ranking.

REFINEMENTS

How is one to proceed when choices are inconsistent or clearly not in an individual's best interests? BR's position is that such choices cannot be ignored. One has to accept that individual choices might not be able to discern preferences sufficiently to be a useful guide to policy makers. At the same time, however, BR are confident that scientific advances in the understanding of psychological and neurological processes will lead to a better understanding of the ancillary conditions that result in poor choices and cloud the welfare analysis. Then, if these ancillary conditions are present when people are making choices, the choices become suspect and the GCSs that include these ancillary conditions can be eliminated. They refer to the process of eliminating ancillary conditions as refining the GCSs. A smaller set of GCSs is more likely to generate a complete ordering of individuals' preferences.

BR offer a number of circumstances that might suggest refinements. They note that individuals may process information incorrectly because they are inattentive to some parts of the constraint set, or they fail to relate their choices to consequences, or still other reasons. With a greater understanding of cognitive processes, we may be able to recognize when such mistakes are likely to be made and ignore the choices made under these conditions for the purposes of welfare analysis. Choices made by habitual users of addictive substances are a second example. Habitual drug use is known to affect users' ability to process information correctly under certain circumstances. A third example relates to the context in which choices are made. Suppose an individual chooses x in (X,d_1), where $d_1 =$ distracted, and chooses y under (X,d_2), where $d_2 =$ focused. The choice of x is suspect enough so that policy makers might properly assume that the choice of y is a better indication of the individual's true preferences. More generally, if people admit that some of their previous choices were mistakes, then these choices are unlikely to reveal their true preferences.

BR are hopeful that such refinements might be possible, but also realistic. They recognize that psychological and neurological processes are often highly complex, such that the ability of researchers to describe general conditions under which people make mistakes or become distracted may be a long way off if, indeed, it is ever to be achieved. They also insist that discarding certain choices as inappropriate should be subject to a very high standard of scientific proof. To casually dismiss choices as inappropriate runs the risk of public officials deciding what is natural for people to believe and they very much want to guard against that happening. Thus they recognize that their framework may not be able to resolve many of the anomalies that behavioral economists have uncovered, at least not in the near future. Nonetheless, mainstream economists are likely to agree with BR that welfare analysis should be a choice based to the greatest extent possible. Simply redefining preferences or utility functions to be consistent with individual anomalies on a case-by-case basis is probably not a useful way to proceed.[14]

REFERENCES

Andreoni, J., February 1995. Warm glow vs. cold prickly: the effects of positive and negative framing on cooperation experiments. Quarterly Journal of Economics 110 (1), 1–22.

Barberis, N., December 2012. Thirty years of prospect theory in economics: a review and assessment. NBER 18621. Working Paper.

Becker, G., 1976. The Economic Approach to Human Behavior. University of Chicago Press, Chicago.

Benartzi, S., Thaler, R., February 1995. Myopic loss aversion and the equity premium puzzle. Quarterly Journal of Economics 110 (1), 73–92.

Bernheim, B.D., Rangel, A., July 2005. Behavioral public economics: welfare and policy analysis with non-standard decision makers. Working Paper NBER 11518.

Bernheim, B.D., Rangel, A., February 2009. Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. Quarterly Journal of Economics 124 (1), 52–104.

Bernheim, B.D., Rangel, A., May 2007. Toward choice-theoretic foundations for behavioral welfare economics. American Economic Association Papers and Proceedings 97 (2), 464–470.

Diamond, P., March 2004. Social security. American Economic Review 94 (1), 1–24.

Fehr, E., Gachter, S., September 2000a. Cooperation and punishment in public goods experiments. American Economic Review 90 (4), 980–994.

Fehr, E., Gachter, S., Summer 2000b. Fairness and retaliation: the economics of reciprocity. Journal of Economic Perspectives 14 (3), 159–181.

Fehr, E., Schmidt, K., 2003. "Theories of fairness and reciprocity—evidence and economic applications,". In: Dewatripont, M., Hansen, E., Turnovsky, S. (Eds.), Advances in Economics and Econometrics, Econometric Society Monographs, Eighth World Congress, vol. 1. Cambridge University Press, Cambridge, pp. 208–257.

Feld, L., Frey, B., 2002. Trust breeds trust: how taxpayers are treated. Economics of Governance 3 (2), 87–99.

Fudenberg, D., September 2006. Advancing beyond advances in behavioral economics. Journal of Economics Literature 44 (3), 694–711.

14. Our presentation of the Bernheim/Rangel model is the barest introduction to their complete model, which appears in Bernheim and Rangel (2009). The QJE paper also contains a lengthy discussion of refinements. A shorter version of their model appears in Bernheim and Rangel (2007). The refinements mentioned in the text are taken from this article, pp. 469–470.

Kahneman, D., Tversky, A., October 1992. Advances in prospect theory: cumulative representation of uncertainty. Journal of Risk and Uncertainty 5 (4), 297–323.

Kahneman, D., Tversky, A., March 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–291.

Laibson, D., 1994. Essays in Hyperbolic Discounting (Ph.D. Dissertation). Economics, MIT Press, Cambridge, MA.

Madrian, B., Shea, D., November 2001. The power of suggestion: inertia in 401(k) participation and savings behavior. Quarterly Journal of Economics 116 (4), 1149–1187.

McCaffery, E., Slemrod, J., 2006. Toward an agenda for behavioral public finance (Chapter 1). In: McCaffery, E., Slemrod, J. (Eds.), Behavioral Public Finance. Russell Sage Foundation, New York.

Mullainathan, S., Schwartzstein, J., Congdon, W., 2012. A reduced-form approach to behavioral public finance. Annual Review of Economics 4, 17.1-1-17.30.

Munnell, A., Sunden, A., March 2006. 401K Plans Are Still Coming up Short. An Issue in Brief, Number 45. Center for Retirement Research, Boston College, Chestnut Hill, MA. Table 1, p. 2.

Pesendorfer, W., September 2006. Behavioral economics comes of age: a review essay on advances in behavioral economics. Journal of Economic Literature 44 (3), 712–721.

Postlewaite, A., 2011. Social norms and preferences (Chapter 2). In: Benhabib, J., Bisin, A., Jackson, M. (Eds.), Handbook for Social Economics. North Holland, Amsterdam.

Rabin, M., 2002. A perspective on psychology and economics. Alfred Marshall Lecture European Economic Review 46, 657–685.

Stigler, George, October 1950. The development of utility theory. II. Journal of Political Economy, 392–396.

Thaler, R., Sunstein, C., 2008. Nudges: Improving Decisions about Health, Wealth, and Happiness. Yale University Press), New Haven, Ct.

Chapter 26

Optimal Federalism: Sorting the Functions of Government within the Fiscal Hierarchy

Chapter Outline

Federalism refers to a hierarchical structure of governments in which each person is, simultaneously, a citizen of more than one government. The United States is an example, with its national government, 50 state governments, and over 89,000 local governmental entities, including cities, towns, counties, regional transportation authorities, metropolitan district commissions, and the like. Each person in the United States falls within the jurisdiction of at least three, and often four or more, distinct governmental bodies. The United States is hardly unique in this regard; all the industrialized market economies have a federalist structure.

A federalist structure adds considerable depth and complexity to normative public sector theory because of its layered jurisdictions. The fundamental principles of public expenditure and tax theory developed in Parts II and III of this book still apply under federalism. In particular, government intervention is still justified by the breakdown of the technical and market assumptions underlying a well-functioning competitive market system, to address such problems as externalities, decreasing-cost production, private information, and market power. In addition, the goal of government intervention remains social welfare maximization, which, broadly speaking, translates into the pursuit of efficiency and equity (as always, stabilization problems will be ignored). As we have discovered, achieving a social welfare maximum is an incredibly difficult task for even a single government. Optimal public sector decision rules are easy enough to describe, but their application is often problematic at best. A federalist structure of governments significantly complicates both the theory and the application of public sector decision rules.

THE POTENTIAL FOR INCOMPATIBILITIES AND DESTRUCTIVE COMPETITION

The complications lie at the heart of a federalist system, that more than one government has jurisdiction over any one person. Given the layered structure, it is all too easy to envision potential inconsistencies and incompatibilities arising if each government simply tries to follow the

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00026-1

single-government decision rules of public sector theory. This is so even if we were to assume that the population is stationary. For example, the national government may want to transfer income from person 1 to person 2, whereas the state government where the two people live may want to do exactly the opposite. Or, one state government may encourage expansion of a decreasing-cost public utility, which pollutes the air over a neighboring state that is trying to reduce air pollution.

People are highly mobile, not stationary, in the developed market economies, and this gives rise to further complications. Mobility has a direct impact on a normative theory of the public sector because people move partly in response to government expenditure and tax policies and then become voters in their new jurisdictions. Hence, their movement can lead to a competition problem for lower level governments. Income redistribution is a common example. If wealthy residents of town A are asked to provide social services to the poor, they may well move to some other town, B, which has no such policy. People's ability to "vote with their feet" forces governments within a given level of the fiscal hierarchy into a competition with one another to attract and retain residents. In general, optimal decision rules must be adjusted as people move in response to them.

Mobility and local competition turn out to introduce another source of potential inefficiency into the economy. They also raise the possibility that no stable equilibrium of localities exists. In short, a federalist structure of governments is unlikely to achieve a social welfare maximum when the population is mobile.

The Two Fundamental Sorting Questions of Fiscal Federalism

At the outset, therefore, federalism poses two fundamental sorting questions that a normative public sector theory must address. The first relates to the allocation of the legitimate functions of government throughout the fiscal hierarchy: Which governments should provide the various legitimate allocational and distributional functions of government so as to avoid potential incompatibilities and destructive competitions among the governments and achieve a social welfare maximum? The second question relates to the sorting of people among jurisdictions: How must people sort themselves among the various jurisdictions, again with the goal of avoiding incompatibilities and intergovernmental competitions and achieving a social welfare maximum? The attempt to answer these two questions is referred to as the *theory of fiscal federalism*.

The two questions of fiscal federalism are naturally interrelated. The sorting of the functions of government among jurisdictions in part determines how people move in response to government policies. In turn, the movement of the people in response to government policies determines in part how the functions should be allocated among jurisdictions.

Social Welfare within Fiscal Federalism

A final point of introduction is that the very meaning of social welfare maximization requires careful attention in a federalist system. A natural extension of the single-government model would be to assume that each autonomous government formulates its own distinct social welfare function that it attempts to maximize. As we have seen, a government has no political identity without a social welfare function in the mainstream normative theory of the public sector. Under this assumption, a natural characterization of an optimal federalism is one in which each government has maximized its own version of social welfare. This is the obvious extension of the standard single-government policy objective to a multigovernment environment and just as obviously is a very difficult objective to achieve.

This is not, however, the usual approach taken in the extensive theoretical literature on the optimal design of a federalist system. Most theoretical models of fiscal federalism assume that only the highest level (i.e., national) government in the fiscal hierarchy has a social welfare function. That is, only the national government concerns itself with the distribution question. The national government may also address allocational problems, but the key point is that all lower level governments (state, local, county, and so forth) concern themselves *only* with allocational problems. This modeling approach to fiscal federalism is a less complicated extension of the single-government model than the assumption that all governments have social welfare functions. But it is a somewhat discomforting framework for a normative theory, since only the national government has a distinct political identity.[1]

In addition, many theoretical models of fiscal federalism employ the first-best technical, market, and policy assumptions to exploit the dichotomization of allocational and distributional issues inherent in first-best (but only first-best) models. The first-best assumptions allow the models to separate the allocational and distributional functions within the fiscal hierarchy. The separation would generally not be possible in a second-best environment.

1. We will wait to pursue this point in detail until the last section of the chapter, in which we analyze distributional issues in the design of an optimal federalist system. Here we will follow the usual modeling approach in the literature.

SORTING THE FUNCTIONS OF GOVERNMENT WITHIN THE FISCAL HIERARCHY

The natural place to begin is with the sorting of functions throughout the fiscal hierarchy because it is the logically prior question. The sorting of people occurs mostly within a single layer of the hierarchy, such as among the localities within a state.

The assumption that only the national government has a social welfare function, when combined with a first-best policy environment, gives rise to a fundamental challenge for the theory of fiscal federalism, namely, what is the advantage of having a federalist structure?

The issue can best be seen as follows. Suppose the national government pursues the norm of social welfare maximization using the traditional first-best analytical framework developed in Chapter 2. In condensed form:

$$\max_{(X_{hi})} W\left[U^h\left(X_{hi}\right)\right]$$

$$s.t.\ F\left(\sum_{h=1}^{H} X_{hi}\right) = 0$$

where $h = 1,\ldots, H$ includes everyone in the society, and $F()$ is the aggregate production−possibilities frontier. If the national government can achieve a set of policies consistent with the first-order conditions of this model in the presence of such problems as externalities, decreasing-cost production, and a nonoptimal distribution of income, what can lower level governments possibly do to enhance the economic well-being of society? Why not let the national government do everything?

Public sector economists have provided a variety of answers to this question, none entirely satisfactory. In considering them, keep in mind that each answer attempts to justify a role for lower level governments only with respect to the standard allocational or efficiency questions. Almost everyone concedes the distributional question to the national government. Social welfare issues are largely absent in lower level or local government decision making in the federalism literature.

Stigler's Prescription for an Optimal Federalism

George Stigler, in his short masterpiece "Tenable Range of Functions of Local Government" prepared for the Congressional Joint Economic Committee, adopted what amounts to an axiomatic resolution of this question Stigler (1957). His justification for local (i.e., lower level) governments rests on two principles.

The first principle is that representative government works best the closer the government is to its constituency (presumably because local governments perceive the utilities or demands of their constituents better than a national government could, although this is unclear from his article). This principle is consistent with the notion that the democratic one-person-one-vote town meeting is the ideal form of government, a notion that has held considerable sway throughout the history of the United States.

The second principle is that subsets of people within a country have the right to vote different kinds and amounts of public services for themselves. This principle is the so-called doctrine of states' rights that was expounded so eloquently by various US southern politicians in the pre-Civil War days (absent, of course, the racial and slavery issues commonly associated with the doctrine during that period of US history). A recent variant of the states' rights doctrine is that allowing for differences in public services encourages healthy experimentation and innovation in the public sector.

The growth in the size and influence of the national government in the United States has diminished somewhat the commitment to these principles. But it is fair to say that they remain persuasive even today, as seen by the current movement to devolve some of the functions that the national government had assumed back to the state and local governments. A recent example is replacement of the Aid to Families with Dependent Children (AFDC) public assistance program with Temporary Aid for Needy Families (TANF) in 1996. TANF gives the states much more discretion in how they choose to assist poor families and make use of the federal funds they receive to support those families.

According to Stigler, these two principles imply that decision making should occur at the lowest level of government consistent with the goals of allocational efficiency and distributional equity. Notice that his conclusion provides, simultaneously, the justification for federalism and the norm for designing an optimal federalist system, one by which the various legitimate functions of government are best allocated among the governments within the fiscal hierarchy. In effect, Stigler has turned our original challenge to federalism on its head by asking: When is it appropriate to have anything but small, local governments?

His answer is that higher level governments may be necessary to achieve either allocational efficiency or distributional equity. In particular, he argues that the national government is the proper government for resolving the distribution question to avoid incompatibilities and competition among governments. As already noted, most other theorists have followed him on this point. In contrast, the responsibility for allocational functions throughout the fiscal hierarchy turns naturally on the geographic scope of both externalities and decreasing costs, the traditional allocational issues in first-best public sector theory. A governmental

body must be sufficiently large to capture all decreasing costs from a particular decreasing-cost service or to include all citizens affected by a particular externality-generating activity, but it need not be any larger. Thus, the optimal size of a jurisdictional unit varies with each specific instance of a decreasing-cost service or an externality.

Oates' Perfect Correspondence

Wallace Oates, in *Fiscal Federalism*, solidified Stigler's principle by proposing the notion of a perfect correspondence[2]:

> *the optimal form of federal government to provide the set of* n *public goods would be one in which there exists a level of government for each subset of the population over which the consumption of a public good is defined. This would be sufficient to internalize the benefits from the provision of each good. Such a structure of government, in which the jurisdiction that determines the level of provision of each public good includes precisely the set of individuals who consume the good, I shall call a case of* perfect correspondence *in the provision of public goods. In the ideal model, each level of government, possessing complete knowledge of the tastes of its constituents and seeking to maximize their welfare, would provide the pareto-efficient level of output and would finance this through benefit pricing.*
>
> *That the allocation of resources resulting from our ideal case of a perfect correspondence is pareto-efficient is, I think, clear (assuming no private sector inefficiencies).*

Given the existence of a federalist system, the notion of a perfect correspondence sets a natural limit on the size of each local government. It is clearly a stringent requirement, leading one to question whether a perfect correspondence for even one public good or decreasing-cost service actually exists, since political boundaries are never determined solely by the extent of externalities or decreasing costs. But a more fundamental theoretical issue turns on the usefulness of perfect correspondence as a policy norm for the public sector. Is it even worth pursuing by restructuring existing jurisdictional boundaries?

Oates is certainly correct when he says that a perfect correspondence generates a first-best social welfare optimum, assuming that local governments follow the first-best allocational decision rules. But we must return to our original challenge posed above. Given a first-best policy environment in which only the national government has a social welfare function, why is local decision making necessary at all, the existence of a perfect correspondence notwithstanding? Why cannot the national government note the extent of each externality or decreasing-cost service and make the appropriate policy response? There is something of an asymmetry here. A nonperfect correspondence can preclude local autonomy, but a perfect correspondence does not necessarily imply local autonomy in order to achieve a social welfare maximum. If we are to make a compelling theoretical argument for a federalist structure, something besides perfect correspondence is required.

Oates' Decentralization Theorem

Oates provides one possible justification by adding a new constraint to the basic first-best general equilibrium model.[3] Following Oates, assume that there are two subgroups of people, A and B, within the total population, such that all individuals within each subgroup have identical preferences but preferences vary across A and B. Suppose, in addition, that society produces two purely private goods, X and Y, that are both consumed by all members of the society. Y happens to be provided by a government, either national or local, despite its being a private good. Assume, finally, that the distribution of income is optimal, so that each subgroup can be viewed as containing a single individual. Under these assumptions, social welfare maximization is equivalent to achieving a pareto optimum, which can be represented as follows:

$$\max_{(X^A,Y^A,X^B,Y^B)} \quad U^A\left(X^A,\ Y^A\right)$$

$$s.t.\ \ U^B\left(X^B,Y^B\right) = \overline{U}$$

$$F\left(X^A + X^B; Y^A + Y^B\right) = 0$$

We know that the first-order conditions for this problem are

$$\mathrm{MRS}^A_{X^A,Y^A} = \mathrm{MRS}^B_{X^B,Y^B} = \mathrm{MRT}_{X,Y} \qquad (26.1)$$

Moreover, with different tastes, $X^A \neq X^B$ and $Y^A \neq Y^B$ in general, at the optimum.

Given the model as it stands, it obviously makes no difference whether a single national government provides

2. Excerpted from *Fiscal Federalism* by Wallace E. Oates, © 1972 by Harcourt Brace Jovanovich, Inc., pp. 34–35 (Oates, 1972). Reprinted by permission of the publisher. Two points are worth noting with respect to Oates' definition of perfect correspondence. First, while he talks only of public goods, the principle clearly applies as well to any form of externality, or any decreasing-cost industry. Second, Oates claims no originality for the notion of perfect correspondence, only for the terminology. Many other authors besides Stigler viewed the ideal federalist structure in a similar vein, including Albert Breton, Mancur Olson, and Vincent Ostron et al. See pp. 34 (note 4) and 35 in *Fiscal Federalism*.

3. Adapted from *Fiscal Federalism* by Wallace E. Oates, © 1972 by Harcourt Brace Jovanovich, Inc., p. 55, by permission of the publisher.

Y^A and Y^B according to Eqn (26.1), or whether each subgroup forms its own government and individually satisfies:

$$\text{MRS}^A_{X^A,Y^A} = \text{MRT}_{X,Y} \tag{26.2}$$

and

$$\text{MRS}^B_{X^B,Y^B} = \text{MRT}_{X,Y} \tag{26.3}$$

Suppose, however, that the national government is constrained to offer equal amounts of Y to each subgroup, so that $Y^A = Y^B$ with national provision of Y. Since, in general, $Y^A \neq Y^B$ at the social welfare optimum, this would represent an additional binding constraint on the formal general equilibrium model, implying a lower level of social welfare at the optimum. It is easy to show that the new first-order conditions become:

$$\text{MRS}^A_{X^A,Y^A} = \text{MRS}^B_{X^B,Y^B} = \text{MRT}_{X,Y} + \frac{\lambda_3}{\lambda_2 F_x} \tag{26.4}$$

where:

λ_2 = the Lagrangian multiplier associated with society's production possibilities, $F() = 0$ and

λ_3 = the Lagrangian multiplier associated with the new constraint, $Y^A = Y^B$.

Local autonomy is obviously the preferred structure under these conditions because it avoids subjecting society to an unnecessary constraint upon government decision making. Oates labels this result *the decentralization theorem*[4]:

> *For a public good—the consumption of which is defined over geographical subsets of the total population, and for which the costs of providing each level of output of the good in each jurisdiction are the same for the central or the respective local government—it will always be more efficient (or at least as efficient) for local governments to provide the pareto-efficient levels of output for their respective jurisdictions than for the central government to provide* any *specified and uniform level of output across all jurisdictions.*

The decentralization theorem does not solve the problem of justifying local level governments in a first-best policy environment. It is really an exercise in the theory of the second best, precisely because the national government is forced to offer equal service levels to all subsets of the population. Nonetheless, this is a compelling restriction in the context of the United States. US citizens have expressed a longstanding fear of standardization if the national government provides public services. There are any number of examples. People have consistently and successfully argued for local autonomy over public elementary and secondary education on the grounds that a federal takeover, despite some financial advantages, would imply standardized education for all children. The Federal Communications Commission has promoted local public television production to offset the standardized sitcom- and sports-dominated programming offered by the national networks. Along these same lines, the national government is prohibited by the Constitution of the United States from varying certain taxes on a geographical basis. The point is that Oates' decentralization theorem strikes a responsive chord, at least in the United States. It is not just some arbitrary formal model that happens to be biased against national decision making.

4. Excerpted from *Fiscal Federalism* by Wallace E. Oates, © 1972 by Harcourt Brace Jovanovich, Inc., p. 35. Reprinted by permission of the publisher.

Misperceived Preferences

Oates' justification for local autonomy is still somewhat unsettling because nationally provided services do not necessarily have to be standardized. A different approach that may be more appealing relies on a particular form of private information. It picks up on Stigler's idea that local officials know best their own constituents' demands for public services.

Suppose that the only allocational problem facing society is the existence of a Samuelsonian public good, X_g, the consumption of which happens to affect only a subset of the population. Let $h = 1,\ldots, k$ be the affected subset and $h = k+1,\ldots, H$ be the unaffected subset. All other goods are pure private goods, and there is no other problem (e.g., decreasing costs) requiring government intervention for allocational reasons. The distribution of income is optimal and determined by the national government.

In a first-best world of perfect certainty, either the national government or a local jurisdiction composed of individuals $h = 1,\ldots, k$ could provide the proper level of X_g in accordance with the standard first-order condition:

$$\sum_{h=1}^{k} \text{MRS}^h_{g,1} = \text{MRT}_{g,1} \tag{26.5}$$

where good 1 is one of the purely private goods. Suppose, however, that the local jurisdiction knows its citizens well in the sense that it knows any individual's $\text{MRS}^h_{g,1}$ with perfect certainty, whereas the national government knows each of these people less well in the sense that it observes each individual's marginal rate of substitution as a random variable:

$$\widehat{\text{MRS}}^h_{g,1} = \text{MRS}^h_{g,1} + \alpha \tag{26.6}$$

where:

$\text{MRS}^h_{g,1}$ = the true MRS as observed by the local jurisdiction and

α = a random variable, with $E(\alpha) = \overline{\alpha}$, possibly 0.

Under these conditions, social welfare is maximized, in general, by having the local jurisdiction form and decide the appropriate level of X_g, rather than letting the national government determine X_g according to the first-order condition:

$$\sum_{h=1}^{k} \widehat{\mathrm{MRS}}_{g,1}^{h} = \mathrm{MRT}_{g,1} \tag{26.7}$$

If $\overline{\alpha} \neq 0$, the national government's decision rule is clearly biased, implying either over- or underprovision of X_g. Even if $\overline{\alpha} = 0$, however, so that $\widehat{\mathrm{MRS}}_{g,1}^{h}$, is an unbiased estimate of $\mathrm{MRS}_{g,1}^{h}$, a risk-averse society would prefer local provision of X_g. Expressed in terms of indirect utility functions:

$$V^h(\overrightarrow{q};\ I^h;\ X_g^*) > E\left[V^h(\overrightarrow{q};\ I^h;\ \overline{X}_g)\right] \quad h = 1,\ \cdots,\ k \tag{26.8}$$

where:

$X_g^* =$ the optimal level of X_g, obtained with local provision and

$\overline{X}_g = X_g^* + \beta$, with $E(\beta) = 0$, obtained with national provision.

Assuming risk aversion, persons $h = 1,\ldots, k$ would be willing to pay a risk premium for local rather than national provision of X_g.

Proponents of federalism probably have this type of uncertainty in mind when they argue that local governments best know the interests of their own citizens. The sheer geographic distance from the central government to most of the people within a given society is bound to affect adversely the transmission of information.

Local Autonomy in a First-Best Environment?

Oates' decentralization theorem and the notion of misperceived preferences justify local autonomy by introducing second-best restrictions—standardization of national services or private information. The question remains whether local autonomy can be justified in a first-best environment when the national government is the only government allowed to make social welfare rankings, and it has perfect knowledge and access to whatever policy tools are necessary to generate first-best allocational decision rules. The answer would appear to be no, yet local autonomy does seem more appropriate for public services that are limited in scope, all the more so when Oates' perfect correspondence happens to obtain within jurisdictions that already exist. Stigler's twin axioms for allocating the functions of government—choose the lowest level jurisdictions consistent with allocational efficiency and preserve states' rights—remain compelling despite the formal implications of first-best theory. Is it possible, therefore, to resurrect fiscal federalism as an optimal governmental structure without introducing specific second-best assumptions? In our view, the answer is "yes": federalism can be justified on distributional grounds, but this involves a line of argument that has not received much attention in the theoretical literature on fiscal federalism.

OPTIMAL FEDERALISM AND THE DISTRIBUTION FUNCTION

The literature on the optimal structure of a federalist system of governments is virtually unanimous in assigning decisions on income distribution to the national government.[5] According to the conventional wisdom, allowing redistribution by lower level ("local") governments in the fiscal hierarchy is formally inconsistent with social welfare optimization, whether one assumes that people are immobile or fully mobile across local jurisdictions.

We happen to disagree with the conventional analysis on this point. In our view, a federalist system is not only formally consistent with social welfare maximization when it contains lower government redistributions, but it also *requires* local redistributions to have meaning as an optimal fiscal system from the mainstream perspective. A review and criticism of the conventional position is useful before developing our preferred model of federalism.

Redistribution, the Competition Problem, and Potential Incompatibilities

Assume first that people are mobile, and suppose that one local government tries to redistribute from its rich to its poor citizens, but only one. Neighboring governments do not attempt any redistribution. The wealthier citizens of the redistributing locality would have an incentive to move to the neighboring localities. This is the competition problem referred to earlier, and it is clearly in evidence in many metropolitan areas in the United States.

Such migration has two unfortunate implications. First, the government that tries to redistribute is totally frustrated. Not only are its poor not made significantly better off, but the total tax base of the community has declined and it becomes more difficult to maintain per capita levels of public services. Second, if people move in response to taxation, it tends to increase the deadweight loss arising

5. A notable exception is (Pauly, 1973). Pauly develops a model based on Hochman and Rodgers' notion of pareto-optimal redistributions (Chapter 10), in which, under certain conditions, local government redistributions are optimal. In this chapter, we argue that local redistribution makes sense for a federalist system even if redistributions are based solely on interpersonal equity considerations without adding an externality component.

from taxation (assuming for the moment that lump-sum redistributions are not viable). Thus, redistributions at the local level are seen to be inconsistent with the goal of maximizing social welfare in a federal system with mobile resources.

The competition problem reaches its full force under perfect mobility, in which people are free to move to any locality and mobility is costless. Fully autonomous local redistribution is impossible in this case since equilibrium requires equal treatment of equals no matter where people live. The fiscal incidence on any one of its citizens is exogenous to each locality.

Even in a world without mobility, incompatibilities can arise throughout a federalist system if more than one government redistributes income. Suppose local government L wants to effect a redistribution from citizens in group A to citizens in group B, but the national government prefers a net redistribution from group B to group A. One can imagine an endless chain of redistributions as each government tries to have its way. Of course, this sort of game must be ruled out, and the most obvious way is to deny one government the right to redistribution.

To avoid the competition problem and potential incompatibilities, therefore, conventional analysis assigns redistribution policy solely to the national government. In an optimal federalist system, all lower level governments in the fiscal hierarchy perform only allocational functions, in accordance with the principles outlined in the preceding section. Furthermore, the prevailing model of optimal federalism stipulates that all local allocational expenditures be financed according to the benefits-received theory of taxation to avoid any unintended redistributions from their allocational decisions. An example would be financing local public goods by Lindahl taxes that equal each person's MRS between the public good and the numeraire good. Only the national government is allowed to tax on some basis other than benefits received, such as ability to pay, and then only to effect the goal of a just distribution. If local governments were to use some tax principle other than the benefits-received principle, then they would likely be redistributing, and the problems of moving to escape taxes, excess burden, and incompatibility among governments are sure to arise. Oates is very clear on the point[6]:

> *The most attractive solution to this whole (distribution) problem (at a formal level at least) is that suggested in Chapter One: let the central government resolve the distribution problem and allow decentralized levels of government to provide public services that they finance with benefit taxes. The use of ability-to-pay taxation by local government, instead of a national negative income tax, may well involve a very high cost both in terms of excess burden and the failure to realize distributional objectives.*

According to Oates, this scheme produces a welfare optimum in an ideal world of perfect correspondence.

Two implications of the conventional model deserve mention. Models of fiscal federalism assume that mobile citizens search for localities offering their most preferred level and mix of public services. Roughly speaking, people choose among localities with high service—high tax, medium service—medium tax, and low service—low tax along a broad spectrum. The public services would only be of the allocational kind, however. Distributional concerns would not enter into their locational decisions because all distributional issues are resolved by the national government. Another implication of the model in the ideal world of perfect correspondence is that there is no need for grants-in-aid among governments. Redistributions occur only among people, and at the instigation of the national government. According to Oates,[7]

> *To achieve a just distribution of income among the individuals in a nation, a national program that redistributes income among individuals, not among jurisdictions, is the preferred alternative.*

Criticisms of the Prevailing Model

To fix ideas on the meaning of a social welfare optimum in a federalist system, we assume a first-best economic and policy environment. This is the appropriate way to assess the conventional position, since it was developed within a first-best context.

In our view, the conventional first-best analysis of optimal federalism is deficient in three respects. It has difficulties with decreasing-cost services, it has questionable political implications, and it flies in the face of reality.

Decreasing-Cost Services

The notion that taxation according to the benefits-received principle necessarily avoids redistributions is not correct, at least not with respect to decreasing-cost services. To preserve efficiency with decreasing-cost services, which an optimal federalist system must surely do, correct benefits-received taxation or pricing implies that price must be set equal to marginal costs. Any other price cannot achieve a social welfare optimum. The problem is that setting price equal to marginal costs is not sufficient to cover full average costs if average costs are declining, so that the local

6. Excerpted from *Fiscal Federalism* by Wallace E. Oates, © 1972 by Harcourt Brace Jovanovich, Inc., p. 150. Reprinted by permission of the publisher.

7. Excerpted from *Fiscal Federalism* by Wallace E. Oates, © 1972 by Harcourt Brace Jovanovich, Inc., p. 81. Reprinted by permission of the publisher.

government has to make up the deficit out of lump-sum taxes and transfers.[8] The question then arises: How is the local government supposed to finance the deficit if, as in the prevailing model, it is constrained from making redistributional decisions? Formally, this restriction implies that it is not allowed to have a social welfare function.

As we saw in Chapter 9, the decision to provide decreasing-cost services in an economy with a single government is inextricably tied to the lump-sum redistributions that satisfy the interpersonal equity conditions of social welfare maximization. The only modification is that the sum of all lump-sum taxes collected from individuals must exceed the sum of all lump-sum transfers to individuals by an amount sufficient to cover all deficits incurred by decreasing-cost industries. In this case, then, allocational and redistributional considerations are also inextricably bound together. A local government cannot, by itself, make what is essentially an allocational decision without simultaneously having some way of ranking individuals, such as by means of a social welfare function, to decide how to finance the deficit. The alternative of reinterpreting the benefits principle of taxation to allow for average cost pricing is clearly illegitimate, because then the system of optimal federalism cannot achieve a welfare optimum. It cannot satisfy the pareto-optimality conditions of first-best theory.

One practical solution to the deficit problem would be to extend the benefits-received principle to the financing of the deficit. Have the local governments institute a two-part tariff, in which consumers pay a price equal to marginal costs to use the service, plus a one-time, lump-sum fee (which potential users would have to pay as well) sufficient to cover the resulting deficit. Believers in the benefits-received principle would be comfortable with this solution, but it is not especially compelling in the mainstream neoclassical model. Recall that the benefits-received principle has no standing as an equity principle in the mainstream model. It can only be applied to public pricing to achieve pareto-optimal allocations, such as setting a price equal to marginal cost for decreasing-cost services. The problem with applying it to the deficit is that it is not distributionally neutral. Therefore, it does not have any particular theoretical appeal if the local government cannot make distributional judgments. Why not have each locality charge just one of its citizens for the entire deficit, with the confidence that the national government's redistribution policies will correct any undue harm suffered by the individuals chosen?

A model of optimal federalism can sidestep the deficit problem by not allowing local governments to make decisions involving decreasing-cost services. These must also be the sole prerogative of the national government. One might counter that the local governments could decide on the level of service to be provided with the national government merely guaranteeing to cover whatever deficit ensues. But whether or not the service is worthwhile depends both upon the demands of the individuals using the service (assume no externalities) and upon the social welfare rankings of these people as determined by the national social welfare function. Since redistributions are the sole prerogative of the national government in the conventional model, the final decision rests in part with that government. Thus, the local governments cannot make a truly autonomous decision in this area if all the tenets of the prevailing model are to be preserved. This is not a devastating blow to that model, merely uncomfortable. Since it excludes decreasing-cost services from complete local autonomy, it probably excludes at least a number of transportation, recreational, and telecommunication services. One thinks immediately of mass transit systems, highways, parks (assuming no congestion), and television and internet cable transmissions. At this point it appears that local governments have only a single decision to make on their own, that of providing services with significant externalities among the local constituents.

Politics and the Social Welfare Function

A second, and more fundamental problem with the conventional solution to the distribution function was mentioned earlier in the chapter: Within the mainstream normative theory of the public sector, in what meaningful sense has an autonomous government been established if that government does not have the ability to determine a set of distributional rankings among its constituents, such as by means of a social welfare function? According to the normative theory, distributional rankings are the only element that the government itself brings to the analysis through a collective political decision; otherwise, it merely accepts consumers' preferences as paramount and acts, in effect, as their agent. Without the distribution function, an autonomous government can hardly be said to exist. The conventional analysis suggests that lower level governments have essentially a single set of decisions to make entirely on their own, those relating to markets with significant externalities among the local constituents. In doing so they merely accept the distribution of income within their jurisdictions as determined by the combination of competitive market forces and national redistribution policies. They are agents pure and simple, sounding out the preferences of their constituents to satisfy conditions such as $\Sigma MRS = MRT$.

8. Whether or not the service is privately or publicly owned is of little consequence. Decreasing-cost industries, if correctly priced, always involve a governmental decision because it is the government that must decide whether the benefit of having the service justifies the cost of financing the deficit.

One begins to wonder why local governments should even bother with externalities. If the national government is engaging in lump-sum redistributions to achieve a just distribution of income in a first-best environment, then it is satisfying a set of first-order interpersonal equity conditions of the form:

$$\frac{\partial W}{\partial U^h}\frac{\partial U^h}{\partial X_{h1}} = \quad \text{all} \quad h = 1, \cdots, H$$

where:

$W = W(U^1 \ldots, U^H) =$ the social welfare function,

$U^1, \ldots, U^H$ is the utility functions of the H individuals in the society, and

$X_{h1} =$ the consumption of good 1 by person h (one can think of good 1 as lump-sum income arising from a fixed factor of production).

But, if the national government knows enough to do this, it certainly knows enough to satisfy the pareto-optimal conditions within each jurisdiction to correct for local externalities. Put differently, if the national government is satisfying the distributional preferences of society, it might as well do everything else. The local governments are clearly not necessary.

We have reached an impasse. On the one hand, local governments have no political input into a formal model of the public sector without social welfare functions. On the other hand, redistributions at lower levels of government dictated by local social welfare functions can generate competition problems or incompatibilities among governments.

One might be tempted to resolve the impasse by permitting all governments to have social welfare functions but allowing only the national government to redistribute lump sum to pursue distributional goals. The problem with this solution is that the notion of a first-best social welfare optimum loses its meaning as a general rule. Consider the situation depicted in Fig. 26.1. Suppose locality L has two people. The curve U^1–U^2 depicts the utility–possibilities frontier for the two people. L_1, L_2, and L_3 are the local government's social welfare indifference curves. Let ray 0C represent an optimal distribution of utility between the two people as determined by the national social welfare function. If forced to be on the ray 0C, the locality will choose point C, but this will not be a first-best optimum from its own citizens' point of view. It is forced into a second-best optimum. If it can redistribute, it will move to D, but then the social welfare function of the national government is not maximized. In either case, it is not clear that society has achieved a welfare optimum, since the citizens belong simultaneously to both governments. Moreover, a compromise solution between C and D on the utility–possibilities frontier obviously satisfies neither government.

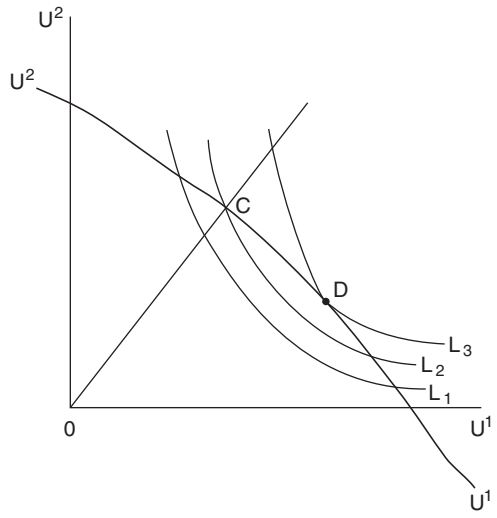

FIGURE 26.1

Redistributions in Reality

Our final criticism of the conventional analysis is simply an appeal to reality. State and local governments in the United States (or any other country) clearly do have distributional preferences. There are any number of examples. State and local governments provide public assistance and other social services to the poor. Questions of choosing among different taxes at all levels of government often consider their perceived incidence. States and localities are concerned about citizen mobility from a distributional perspective. States worry that increases in their public assistance payments will encourage in-migration of the poor from other states. High-income communities use zoning laws in the form of minimum lot sizes to prevent entry of low-income households. These examples all imply that states and localities make social welfare rankings.

It is also true that social welfare rankings differ among localities, states, and the national government. In general, the citizens in any given lower level government do not simply accept the national social welfare ranking as necessarily just, an assumption crucial to the conventional model. Furthermore, these differences in distributional preferences can be given a broader interpretation. People choose different jurisdictions not only because they demand different kinds of public allocational type services but also because they choose to live with people whom they deem compatible in terms of such factors as education, cultural background, and so forth. To deny the latter point is to deny an important justification for a federal system of governments. In essence, federalism supports fraternalism, the principle of states' rights applied to the distribution question in its broadest sense. We are not suggesting that people should pay heed to these factors. The ability to isolate oneself from "undesirables," which federalism

permits, may itself be viewed as undesirable. If one thought so, this would be a strong argument against establishing autonomous local governments. Federalism is not necessarily an optimal form of government.

The Need for Local Social Welfare Functions

In conclusion, we would argue that an optimal structure of fiscal federalism within the traditional theory of the public sector requires a schema whereby each government can simultaneously maximize its own social welfare function, subject to the usual generalized production constraints and market clearance. This is so for two reasons:

1. A truly autonomous government does not exist within the traditional normative theory of the public sector unless it has a social welfare function or some such means of deciding the relative ethical rankings of its constituents.
2. In a federalist system of governments, social welfare maximization by each government as defined above is the only acceptable meaning of an overall first-best social optimum.

Expanding on the second point, recall that the central theoretical problem in designing an optimal federalism is to divide the functions of the public sector among the governments so as to retain the maximum degree of local autonomy while avoiding conflicting decisions among the governments. One must accept the fact that people simultaneously pay allegiance to more than one government and that inconsistencies are almost certain to arise. One manifestation of this point already referred to is that governments will have different social welfare rankings, in general. Given this problem, the suggested definition of a first-best social welfare optimum is the only apparent possibility.

Optimal Redistribution in a Federalist System: An Alternative Model

The basic ingredients of our search for a formal model of federalism were presented in the preceding section. We seek a model in which:

1. Each government simultaneously maximizes an individualistic social welfare function subject only to generalized production constraints and market clearance. This serves as the definition of a first-best social welfare optimum in a federalist system.
2. Autonomy in the decision-making process is preserved at the lowest possible level of government. Without this assumption, the motivation for developing a federalist system effectively collapses. One can always describe a model in which the national government does everything.

Assume a perfect correspondence so that externalities (and scale economies) are entirely contained (exhausted) within each jurisdiction.

A comment is in order before presenting the model. Potential incompatibility is a central feature of a federalist system because citizens are simultaneously members of more than one government. If the first condition is satisfied, these inherent incompatibilities will have been dealt with in a particular way; they will not have disappeared. Therefore, the theoretical problem of determining an optimal federalist system can be restated as follows: What minimum restrictions must be placed on a federalist system of governments to ensure both of the above conditions? Clearly some restrictions must be placed on at least some governments. Incompatibilities are almost certain to arise if each government has the standard Bergson–Samuelson social welfare function used in single-government models. Each government cannot have social welfare functions whose arguments are the utility functions of their individual constituents.

The prevailing model places the restriction, unacceptable in our view, that no government but the national government can have a social welfare function. Our alternative model can be thought of as one with more acceptable restrictions.

In our opinion, the model that is consistent with federalism requires a dynastic set of social welfare functions, as follows: Each government has an individualistic social welfare function whose arguments are the social welfare functions of the governments *immediately* below it in the hierarchy of governments. The lowest level governments have individualistic social welfare functions whose arguments are the utility functions of their constituents—that is, the standard Bergson–Samuelson social welfare function.

In terms of the United States, the national government's social welfare function would contain as arguments the social welfare functions of the 50 states, each state's social welfare function would have as arguments the social welfare functions of the localities within the state, and each locality would have a social welfare function with the utility functions of its constituents as arguments.

To simplify notation, consider a two-tiered federalist system with a national government and L local governments.

Let:

$U^{h1}(X_k^{h1})$ be the utility function of person h living in locality 1.

$h = 1,\ldots, H.$

$1 = 1,\ldots, L.$

with X_k^{h1} the kth good consumed by a person h living in locality 1, for $k = 1,\ldots, N.$

(Note: there are only H people. People are double subscripted according to who they are and where they live.)

Also, let:

$L^1[U^{h1}(X_k^{h1})]$ be the social welfare function of locality 1, whose arguments contain the utility functions of all persons (or potential persons) living in locality 1 and

$F[L^1(U^{h1}(X_k^{h1}))]$ be the national social welfare function with L arguments, $L^1,\ldots, L^L$.

The restrictions on this model consist of the arguments that are allowed to appear in each government's social welfare function.

In this model, allocational decisions are determined exactly as in the prevailing model. The local governments make all decisions on services exhibiting economies of scale and/or externalities as long as the extent of the externalities or scale economies is contained within the local jurisdiction. The national government would provide those services with spillovers across localities, or design grants-in-aid to ensure efficient solutions at the local level (to be discussed in Chapter 28). Each government would maximize its own social welfare function subject to resource and generalized production constraints and market clearance. The usual first-best pareto-optimal conditions would emerge in each case because any social welfare terms drop out from this set of first-order conditions.

The difference with respect to the conventional model is that every government would also be engaged in lump-sum redistributions to satisfy the interpersonal equity conditions. Let good 1 be the good transferred lump sum. The first local government must satisfy the following relationships:

$$\frac{\partial L^1}{\partial U^{h1}}\frac{\partial U^{h1}}{\partial X_1^{h1}} = \quad \text{all} \quad h \text{ in } 1, \text{ every } 1 = 1,\cdots,L$$

The national government satisfies the following interpersonal equity conditions:

$$\frac{\partial F}{\partial L^1}\frac{\partial L^1}{\partial U^{h1}}\frac{\partial U^{h1}}{\partial X_1^{h1}} = \quad \text{all} \quad h = 1,\cdots, H$$

Notice, however, that the redistributions of the local governments ensure that the last two terms of the expression are equal for all people within a given locality, l. Therefore, all the national government needs to do is tax and transfer income lump sum among localities until the entire term is equal for all people. At that point, its social welfare is also maximized.

As an example consider two localities, 1 and 2. By their actions

$$\frac{\partial L^1}{\partial U^{h1}}\frac{\partial U^{h1}}{\partial X_1^{h1}} = \quad \text{all} \quad h \text{ in } 1$$

and

$$\frac{\partial L^2}{\partial U^{h2}}\frac{\partial U^{h2}}{\partial X_1^{h2}} = \quad \text{all} \quad h \text{ in } 2$$

If $\partial F/\partial L^1() > \partial F/\partial L^2()$, the national government would transfer income from 2 to 1 (and the localities would redistribute to maintain social equality on the margin within each jurisdiction). Presumably, the marginal social utility of income of the citizens of 1 would drop and that of 2 would rise (all from the national viewpoint). Redistribution continues until:

$$\frac{\partial F}{\partial L^1}() = \frac{\partial F}{\partial L^2}() \tag{26.9}$$

The same schema holds for an n-tiered hierarchy of governments.

Comments on Our Alternative Model

The advantages of this model over the conventional model of optimal federalism are twofold. First, each government has an identity as traditionally defined in the theory of the public sector, that is, each government is allowed a social welfare function. Consequently, all governments provide important inputs into policy decisions and each retains the ability for truly autonomous decision making over the standard microeconomic functions assigned to the public sector. Second, the definition of a first-best social welfare optimum in a federalist system has been clarified and is consistent with the traditional definition of a first-best social welfare optimum with a single government. Both pareto optimality and interpersonal equity conditions in terms of individuals are satisfied at all levels of government.

The major *operational* difference between the two models is that grants-in-aid among governments now play a central role, even if there exists a perfect correspondence for allocational functions. It is no longer true that redistributions among people at the national level are the "preferred alternative," as Oates claimed. In the alternative model presented here, only the lowest level governments redistribute among people. The higher governments use grants-in-aid to other governments exclusively in their redistributions.

The United States recognizes both models in its redistribution policies and cannot seem to decide which is the better approach to the distribution question. On the one hand, there are a number of national transfer programs, such as Social Security, SNAP (Supplemental Nutrition Assistance Program-"Food Stamps"), and the EITC (Earned Income Tax Credit), that transfer income directly to people. On the other hand, public assistance (welfare) was strictly a state and local initiative until the Great Depression forced the federal government to become involved. Despite the entry of the federal government,

major elements of the US public assistance effort remained essentially state programs. In particular, the states determined the level of monthly payments for the poor who qualified for assistance under the programs. The role of the federal government was primarily to offer financial assistance to the states, with the federal share of the costs dependent in part upon the relative fiscal capacities of the state governments. The replacement of AFDC with TANF further increased state autonomy in providing for impoverished families with dependent children. Public assistance, therefore, has always been structured in line with our alternative model. The United States cannot seem to decide which model for resolving the distribution question is the better approach.

Finally, notice that our alternative model avoids the two problems that proponents of the conventional model perceive as potentially devastating to the federalist system if lower level governments are allowed to redistribute income. Our alternative model obviously avoids the incompatibility problem with nonmobile populations, given the permissible arguments of each government's social welfare function. It also, at least formally, avoids the competition problem with mobile populations. Mobility of the kind that plagues US cities today is a problem partly because the rich who leave do not adequately compensate those remaining behind for the loss in resources when they move out. The US commitment to federalism (that is, to autonomous local governments) supports this phenomenon, which certainly contributes to inequality of opportunity in this country. A number of state supreme courts have questioned the legitimacy of local autonomy in ruling that financing education primarily through local property taxes is inherently discriminatory.

The conventional model suggests that the answer to unwanted inequality lies in stronger national redistributive policies. This may well work, but it represents a movement away from the federalist system. The alternative model presented here suggests an approach that would strengthen the federalist system. If the wealthy residents of city A move to suburb B because city A decides to redistribute income to its poor, presumably the state will insist upon a redistribution from B to A in order to maximize its own social welfare function. Upon knowing that such compensation is required, the incentive to move would diminish. Should the state fail to redistribute in this way, city A is the clear loser, but this is a matter of the state's preferences, not a formal inadequacy of our alternative model. If, as a practical matter, lower level governments within the federal hierarchy are seen to be acting perversely, then one would not want a federalist system in which lower level governments make truly autonomous decisions. There is certainly nothing sacred about a federalist system of governments. We have only suggested that our alternative model is consistent with the notion of a first-best social welfare optimum given the existence of a federalist system.

A final comment is that nothing can preserve complete local autonomy in a world of perfect mobility. Horizontal equity—equal treatment of equals—is the only possible equilibrium condition under perfect mobility no matter how society tries to structure its redistributive responsibilities. Therefore, all governments must accept the same degree of vertical equity—of inequality—throughout the nation. Nonetheless, permitting local social welfare functions gives localities a say in determining how much inequality a society will allow. We will return to the effect of mobility on local redistributions in Chapter 27.

REFERENCES

Oates, W., 1972. Fiscal Federalism. Harcourt Brace Jovanovich, New York.

Pauly, M., February 1973. Income redistribution as a local public good. Journal of Public Economics 2 (1), 35–58.

Stigler, G., 1957. Tenable range of functions of local government. In: Federal Expenditure Policy for Economic Growth and Stability. Joint Economic Committee, Subcommittee on Fiscal Policy, Washington, D.C.

Chapter 27

Optimal Federalism: The Sorting of People within the Fiscal Hierarchy

Chapter Outline

For a given distribution of the population throughout a nation, it is a reasonably simple exercise to define various examples of externalities and decreasing costs over subsets of the entire population and then describe an optimal set of local jurisdictions that can correct these problems in an optimal manner. But there remains the important question of whether people will naturally group into subsets congruent with the set of local jurisdictions required for a social welfare optimum.

Charles Tiebout,[1] the founding father of the mobility literature, was optimistic about federalism. He conjectured that the jurisdictions would form as required. Tiebout argued that the great advantage of federalism compared with having a single government was that it permitted individuals to "vote

1. Tiebout (1956). Tiebout's article is the seminal work, the first to consider the gains from local jurisdictions in a neoclassical framework.

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00027-3

with their feet," as they search for the combination of local services and taxes that maximizes their utility. Tiebout believed that if all people were free to search in this fashion, and packages of services and taxes were replicable, then social welfare would be maximized. This was so for two reasons. First, the ability to search for one's most preferred level of public goods avoids the free-rider problem associated with nonexclusive goods in the single-government model. People naturally reveal their preferences as they search among localities. Second, people with the same tastes will congregate together,[2] thereby providing a better match of preferences to the level of public services provided. Also, the public services will be offered at minimum cost. No cost differences can persist across localities offering identical services because people will naturally gravitate from high-cost to low-cost towns. In effect, the market for local public services will be perfectly competitive.

Tiebout spawned a huge literature that tested his conjecture using formal models, both positive and normative analysis. The positive analysis considers how people sort themselves among the localities. The normative analysis judges the outcomes of the sorting process using the standard efficiency and equity norms.

The literature has generally not supported Tiebout's conjecture; the problem of forming optimal jurisdictions turns out to be much more subtle than Tiebout had imagined. The positive analysis has shown that the sorting process may not reach an equilibrium—some people always want to move to another locality. Normative judgments are moot absent an equilibrium. Furthermore, even if the sorting process does reach an equilibrium, the outcome is often not optimal. The ability of people to move in response to government policies introduces possibilities for inefficiency even though it may produce a better match of preferences for the local public services. Tiebout's conjecture that federalism produces a social welfare optimum obtains only under highly specialized conditions that are unlikely to hold in most practical settings.

The literature on mobility following Tiebout is among the largest in all of public sector economics, so large that we cannot hope to do it justice here. Our more modest goal is to highlight some of the principal modeling techniques and results in the literature.

THE MODELING DIMENSIONS

Models of mobility under federalism vary along at least eight dimensions that influence the results predicted by the model. The dimensions include the underlying economic environment, the nature of the local government sector, and the information set available to citizens within a locality. The following list captures the main distinctions among the models in the literature, although by no means the only distinctions.

The Underlying Economic Environment

Flexible or Fixed Number of Communities

Models that assume a flexible number of communities typically envision people settling a new frontier that was previously uninhabited. Communities form and provide public services. If people do not like the outcome, they can join with other dissatisfied people and form another community offering a different mix of services. Communities continually form and break apart as people search for an equilibrium. The fixed-community models apply to more developed nations. In one variation, the number of communities is fixed but not their size. In another variation, both the number and the size of the communities are fixed. There are a given number of housing sites across all communities that just equals the total population, and equilibrium requires that people sort themselves among the existing sites such that no one wants to move again.

Endowment Income or Earned Income

Some models assume that people are endowed with a given amount of income that they bring with them as they move from one community to another. Some of the income is taxed to pay for the public services. Other models assume that the private and public goods have to be produced within each community, so that income is earned as a result of the production. The factors of production may be labor, land, and capital; just labor and land; or just labor. If just labor, there may be only one kind of homogenous labor or two classes of labor with different skill levels. Also, the output/income from production may be uncertain because of random shocks to the production function. The shocks may be favorable or adverse and either national in scope or idiosyncratic to localities. The income earned from production, or some portion of it, may be taxed to pay for the public services.

The Housing Market

The nature of the housing market is tied to the choice of flexible or fixed communities. The market for land is irrelevant in the frontier models because land is assumed to be available in unlimited amounts at a fixed or no charge. The housing services simply become part of the composite commodity. In contrast, the housing market can become a central feature of a fixed community model, especially if the number of housing sites is fixed. A housing market also allows for the possibility of financing the local public services with property taxes. As a general rule, the operation

2. As George Stigler put it, people would choose among high service—high tax, medium service—medium tax, and low service—low tax communities. See Stigler (1957).

of housing markets prevents an economy of fixed communities from reaching a first-best optimum. This is especially so in models with property taxes, since the property tax itself is a distorting tax.

The Local Government Sector

The Government's Objective Function

There is quite a bit of variety here depending on whether the government officials are utility or profit driven. A natural objective in a normative analysis is to achieve a social welfare optimum, or at least a pareto optimum if the local governments are denied social welfare functions, as they most often are. Profit-driven models typically take one of two forms. In one version, the community is controlled by local developers whose goal is to attract citizens so as to maximize their profits. In another version, the community is controlled by one subset of citizens, immobile landlords who own all the land (housing sites). The landlords try to attract the mobile subset of the population who are searching among communities and who pay rent to the landlords in their chosen community. The landlords (developers) offer public services with the goal of maximizing the rent (profit) they receive.

The Political System

The assumed political systems vary every bit as much as the governments' objective functions. One popular choice is voting for public services by direct democracy—the town meeting model—along with the assumption that the preferences of the median voter are decisive. The median voter is the one whose preferences for the public services lie at the midpoint of the distribution of preferences among all the members of the community. Other models assume a representative two-party system in which the majority party prevails. In models with profit- and rent-maximizing developers and landlords, the developers and landlords are usually assumed to have complete control over the public service and tax policies, although they have to pay attention to the preferences of the mobile citizens they are trying to attract.

The Public Services

The public service is usually a Samuelsonian nonexclusive good within the locality, that is, its services are available to all residents of the locality but not at all available to nonresidents. A common modification is that the good may be subject to congestion. Congestion means that the amount of the good's services available to each person diminishes as the population increases. If the amount of the services per person diminishes in direct proportion to the population, then the public good has the same attributes as a private good. Hence, the congestion feature permits a specification of the public service that varies along the full spectrum from a nonexclusive good to a private good. One important result in the literature is that congestion of some form is necessary to justify local governments when nonexclusive goods are the only activity requiring public sector intervention.

Taxes

The most common choices are a lump-sum head tax, a property tax, and various kinds of income taxes, such as taxes on wage income or rents (profits). Not surprisingly, lump-sum taxes are often required for efficient outcomes.

The Knowledge Set

Nash or Other Behavior

The main distinction here is how savvy the individuals are within each community. Do they take the policies of other communities as given as they make their own decisions about public services and taxes? Or, do they assume that people in other communities react to their decisions in a utility-maximizing fashion? The distinction matters because decisions of any one community generate externalities for all other communities as people move in response to the decisions. As expected, federalism is more likely to achieve efficient outcomes if the mobility externalities are internalized. Note, also, that assumptions about people's reactions to policies in other communities are relevant only in the fixed-community models.

JURISDICTION FORMATION IN ACCORDANCE WITH THE THEORY OF CLUBS

The natural place to begin is with a model of mobility that generates a social welfare optimum, in line with Tiebout's conjecture. The assumption of flexible communities is the one most compatible with Tiebout's thinking, a frontier environment in which communities can form, break apart, and reform to generate the public service levels that subsets of people most prefer. The housing market is irrelevant in such a market, as are information assumptions.

Flexible-community models ask three interrelated questions:

1. Are there incentives for the formulation of local jurisdictions to provide traditional public services such as Samuelsonian nonexclusive public goods?
2. Will the resulting local public services be provided in accordance with standard first-best decision rules, such as Σ MRS = MRT?
3. Will jurisdictions form in such a manner that the public service is provided at least cost?

If the answer to all these questions is "yes," then the outcome can be a social welfare maximum with some additional assumptions.

The models used to analyze these questions draw heavily on Buchanan's theory of clubs (Buchanan, 1965). Briefly, Buchanan argued that determining the optimal membership of any club has an externality element to it. Think of a swim club. On the one hand, accepting new members reduces the direct out-of-pocket costs to the current members by spreading the costs associated with the swimming pool and clubhouse over more people. On the other hand, the new members generate external diseconomies in the form of a more crowded pool. Thus, the optimum-sized membership occurs when the marginal costs of the external diseconomies just equal the marginal savings from spreading total operating costs. A related issue is the optimal size of the pool for a given membership.

The theory of optimal clubs can be adapted quite easily to explain the optimal formation of local jurisdictions along with the provision of local public services. It can also be used to justify the existence of local jurisdictions. We will consider a simple model that Martin McGuire used to analyze this problem (McGuire, 1974).

To fix ideas, begin with a baseline model of a nonexclusive good that is consistent with the model in Chapter 6, a model that does not have clublike features. Suppose a country consists of H identical people whose preferences are defined over two goods, X, and Y^h, where

$X =$ a Samuelsonian nonexclusive public good provided by a government.
$Y^h =$ the income of person h assumed fixed (alternatively, an endowment of a composite commodity with $P_y \equiv 1$).

Preferences are given by

$$U^h\left(X, Y^h\right) \quad \text{all } h = 1, \ldots, H \tag{27.1}$$

Rather than defining a production function relating X and the Y^h, assume first-best production efficiency and posit a cost function for X:

$$C = C(X; \text{other arguments}) \tag{27.2}$$

where C is measured in dollars, the same as the Y^h. If we assume that

1. Income is optimally distributed,
2. $C = C(X)$, with no other arguments, and
3. The costs of X are shared equally by all people by means of head taxes,

then this representation of the Samuelsonian public good is equivalent to the formulation in Chapter 6.[3] To see this, note that the utility of each person h with equal cost sharing is

$$U^h\left[X, Y^h - \frac{C(X)}{H}\right] \quad h = 1, \ldots, H \tag{27.3}$$

3. The assumption of equal cost sharing is convenient but unnecessary, as long as the cost sharing is lump sum.

Since all people are identical and the distribution of income is optimal, all the government need do is maximize Eqn (27.3) with respect to X. The first-order conditions are

$$\frac{\partial U^h}{\partial X} - \frac{\partial U^h}{\partial y^h} \cdot \frac{\partial C}{\partial X} \cdot \frac{1}{H} = 0 \tag{27.4}$$

where

$$y^h = \left[Y^h - \frac{C(X)}{H}\right] \equiv \text{disposable income}$$

Rearranging terms,

$$H \cdot \frac{\frac{\partial U^h}{\partial X}}{\frac{\partial U^h}{\partial y^h}} = \frac{\partial C}{\partial X} \tag{27.5}$$

or $H \cdot \text{MRS}_{X,y^h} = \text{MCX} = \text{MRT}_{X,y^h}$, with $P_y \equiv 1$. Equation (27.5) is the familiar first-best decision rule for public goods, implying national provision of X to all people within the country.

X must have two properties for local provision of the public good to be optimal and analogous to a club: excludability and congestion. Excludability means that if some locality provides an amount of X to its constituents, it can effectively prevent all other people from consuming its X. In terms of the theory of clubs, the services of the club are nonexclusive to its own members, but excludable to nonmembers (i.e., only swim-club members can use the pool).

Excludability alone is not sufficient for optimal provision of X at the local level. The second requirement is that the good must be subject to congestion. As X is provided to more and more people, each person receiving X bears increased costs in some form.

The additional costs can be modeled in one of two ways. It may be that each person's enjoyment of X diminishes as more people consume it, along the lines of a straight consumer externality. This assumption implies that utility is a function of X, Y^h, and N, where N is the number of people consuming X, with $U_N < 0$. Alternatively, the direct costs of providing X could vary directly with N, so that $C = C(X, N)$, with $\partial C/\partial N = C_N > 0$.

With each person bearing some of the direct cost of providing X, it hardly matters in a formal sense which method is chosen. One can think of the cost function as including the external diseconomies of crowding, so that the two stories are virtually identical. All that matters is that each person's utility depends inversely upon N, the number of people consuming the public good. Examples might include police protection and education, in which the quantity of X commingles with certain quality attributes that vary with N to determine the cost of providing a unit of service. For example, police services can be replicated as more people move into a district, but the sheer increase in numbers may cause the costs of controlling criminal activity to increase more than proportionately.

In fact, McGuire chooses the direct cost approach, writing

$$C = C(X, N) \quad C_X, C_N > 0 \tag{27.6}$$

With equal sharing of the costs, the utility of person h becomes $U^h[X, Y^h - C(X, N)/N]$. That is, each person pays the average costs of X where the average is defined relative to N for a given X. McGuire further assumes that the average costs are U-shaped, as depicted in Fig. 27.1. The spreading effects of having N in the denominator dominate up to some point, after which the marginal crowding costs (C_N) dominate, causing average cost to increase.

Under the twin assumptions of excludability and congestion, society has to determine the optimal provision of the good within each jurisdiction and the optimal size of each jurisdiction. Formally, society's problem becomes

$$\max_{(X,N)} U^h\left[X, Y^h - \frac{C(X,N)}{N}\right]$$

The first-order conditions are

$$X(\text{optimal provision}): \frac{\partial U^h}{\partial X} - \frac{\partial U^h}{\partial y^h} C_X \frac{1}{N} = 0 \tag{27.7}$$

$$N(\text{optimal size}): \frac{\partial U}{\partial y^h}\left(\frac{-NC_N + C}{N^2}\right) = 0 \tag{27.8}$$

Rearranging terms,

$$X : N \cdot \left(\frac{\frac{\partial U}{\partial X}}{\frac{\partial U}{\partial y^h}}\right) = C_X \tag{27.9}$$

$$N : \frac{C}{N} = C_N \tag{27.10}$$

Notice that Eqns (27.9) and (27.10) are both functions of X and N, so that the provision of X within each jurisdiction and the optimal size of each jurisdiction are determined simultaneously. Nonetheless, each equation separately has a familiar interpretation. Equation (27.9) says that, given N, each jurisdiction should follow the usual public good decision rule, Σ MRS = MRT, to determine the optimal amount (or size) of X. Equation (27.10) says that, given X, people should form groups such that the average cost of X just equals the marginal costs of one additional person. This is the minimum efficient scale of operation, the long-run AC_{min} that occurs in competitive markets. As long as jurisdictions can be replicated, people can always regroup until AC_{min} obtains in each jurisdiction. No one should have to bear average costs higher than the minimum. Furthermore, Eqns (27.9) and (27.10) imply equal-sized jurisdictions with identical people, H/N in number. National provision of a single X to all H people is no longer optimal, as long as AC_{min} occurs at $N^* < H$.[4]

McGuire also considers the case of a heterogeneous population consisting of homogeneous subgroupings. Without reproducing that model, it is intuitive that each jurisdiction consists of people with like tastes, and that conditions (27.9) and (27.10) hold within each jurisdiction. X has to be provided to people of like tastes in order to maximize each person's net benefit of consuming and paying for X. The only substantive difference is that the level of X varies across jurisdictions, depending on tastes.

Curiously enough, although the McGuire model talks about the simultaneous problems of providing public goods and forming local jurisdictions, it does not necessarily imply local autonomy. The national government could still be the sole supplier of X. It would simply note that the costs of X vary with access to X, so that it would not be optimal to provide a national level of X with access to all, but rather exclusive subsets of X in accordance with conditions (27.9) and (27.10). With homogeneous populations, the amount of X provided to each subgroup would be equal, but these amounts would differ from the single amount of X provided if access did not affect direct costs or create external diseconomies. All McGuire has really done is complicate the nature of the production of X (or the externality associated with X). The national government could, in principle, anticipate this complication even though it is difficult to imagine national provision of local services such as police.

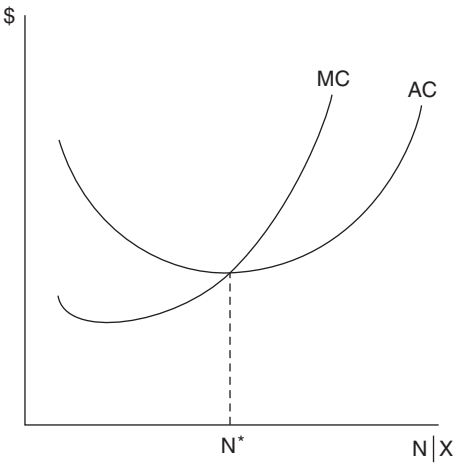

FIGURE 27.1

4. The AC_m in solution can be thought of as an instance of perfect correspondence even though Oates defined a perfect correspondence with respect to an externality that affected a distinct subset of people (firms). Here the additional costs associated with N vary continuously with N. Nonetheless, one can consider X as having two attributes, an externality associated with the public good quality of X and a decreasing cost element associated with the relationship of costs to N. For this good, a perfect correspondence occurs when the decreasing costs are exhausted. Since the X are exclusive to each jurisdiction, the externality associated with X automatically satisfies the perfect correspondence criterion once the jurisdictions have been set. Another point to note is that H/N^* may not be an integer. This possibility causes minor technical problems that need not concern us. For a full analysis, see Scotchmer (1994).

Furthermore, local autonomy cannot guarantee by itself that the optimal conditions will obtain because people may not actually form subgroups in an optimal manner. Indeed, McGuire was forced to impose the following rather complex scenario to ensure that the pareto-optimal conditions obtain in the final equilibrium.

Reaching the Optimum

Imagine, at first, a single individual searching among local jurisdictions that have already been established, but only on a temporary basis. Suppose the individual currently belongs to "temporary" community j. In deciding whether to move from community j to some other community, k, the individual compares the benefits of the move with its costs. The benefits arise from the difference in the public goods provided in each community, X^k versus X^j. The costs depend upon the payment or tax scheme used by each town.

According to McGuire, a natural assumption is that each town asks a new member to pay the marginal costs of entry, so that no existing town member loses by having a new entrant. This is a standard assumption of mobility models. Thus, the cost comparison is $C_N^k(X_k, N_k)$ versus $C_N^j(X_j,\ N_j)$, where C_N is the marginal cost of X in terms of N.

Consider a move to a marginally different community. The change in costs can be represented as the total derivative of C_N:

$$\mathrm{d}C_N = C_{NX}\mathrm{d}X + C_{NN}\mathrm{d}N \tag{27.11}$$

Recall that N and X are simultaneously determined, so that a change in N changes the optimal level of X. Dividing Eqn (27.11) by dX defines the marginal cost/benefit ratio that is available to the individual if he or she moves:

$$\frac{\mathrm{d}C_N}{\mathrm{d}X} = C_{NX} + C_{NN}\left.\frac{\mathrm{d}N}{\mathrm{d}X}\right|_{\text{supply}} \tag{27.12}$$

Equation (27.12) is the individual's ability to trade private goods for the public good on the margin.

Consider, next, the individual's preferences for such a move. The marginal rate of substitution between X and y^h,

$$\mathrm{MRS}^h_{X,y^h} = \frac{\mathrm{d}Y^h}{\mathrm{d}X} = -\frac{\mathrm{d}C_N}{\mathrm{d}X}$$

indicates the individual's willingness to trade private goods for the public good on the margin (an increase in costs subtracts, dollar for dollar, from private income Y^h). Hence, the individual searches until the willingness to trade equals the ability to trade, or

$$\mathrm{MRS}^h_{X,y^h} = C_{NX} + C_{NN}\left.\frac{\mathrm{d}N}{\mathrm{d}X}\right|_{\text{supply}} \tag{27.13}$$

Homogeneous groupings naturally form if many people are searching under these conditions, since localities offering a given marginal cost/benefit ratio ultimately attract only those people whose MRS_{X,y^h} equals that ratio (assuming sufficient ability to form towns so that everyone is in equilibrium). Hence,

$$N\cdot\mathrm{MRS}_{X,y^h} = NC_{NX} + NC_{NN}\left.\frac{\mathrm{d}N}{\mathrm{d}X}\right|_{\text{supply}} \tag{27.14}$$

must hold in the final equilibrium, where N represents the number of people in a particular homogeneous subgroup i.

McGuire argues next that the optimality conditions (Eqn (27.10)) ($C/N = C_N$) must necessarily hold in the final equilibrium if localities can be replicated sufficiently. With each person paying the marginal costs of entry, C_N, the only way that total tax payments can equal the total costs of providing X is if MC equals AC, or $C_N = C/N$. If C_N were temporarily in excess of C/N in some towns and below C/N in others, all of which offer equal levels of X, the people in the high-cost towns would move to the low-cost communities until all profits or rents to existing members disappear. In this sense, the search acts as a competitive market mechanism.

The McGuire search procedure, then, establishes two equilibrium conditions:

$$\frac{C}{N} = C_N \tag{27.15}$$

and

$$N\cdot\mathrm{MRS}_{X,y^h} = NC_{NX} + NC_{NN}\left.\frac{\mathrm{d}N}{\mathrm{d}X}\right|_{\text{supply}} \tag{27.16}$$

Equation (27.15) is one of the two conditions for a welfare optimum. It remains to show that Eqns (27.15) and (27.16) together imply the second pareto-optimal condition, the standard public goods decision rule (Eqn (27.9)).

To see that they do imply Eqn (27.9), totally differentiate Eqn (27.15) to obtain

$$C_X\mathrm{d}X + C_N\mathrm{d}N = NC_{NX}\mathrm{d}X + NC_{NN}\mathrm{d}N + C_N\mathrm{d}N \tag{27.17}$$

Rearranging terms,

$$(C_X - NC_{NX})\mathrm{d}X = NC_{NN}\mathrm{d}N \tag{27.18}$$

$$\frac{(C_X - NC_{NX})}{NC_{NN}} = \left.\frac{\mathrm{d}N}{\mathrm{d}X}\right|_{\text{supply}} \tag{27.19}$$

Substituting Eqn (27.19) into Eqn (27.16) and simplifying yields,

$$N\cdot\mathrm{MRS}_{X,y^h} = C_X \tag{27.20}$$

as required. Tiebout's conjecture holds true in McGuire's fluid, frontier model of fiscal federalism.

FIXED COMMUNITIES AND HOUSING SITES: ADDING THE HOUSING MARKET

McGuire's frontier model can reasonably ignore the housing market on the grounds that supply of land is perfectly elastic in frontier regions where jurisdictions are forming, breaking apart, and reforming, with each town replicating all others in the final equilibrium. The housing market cannot be ignored in models with a fixed number of jurisdictions, however, because then property values are necessarily tied to the provision of public services. Suppose some town offers a particularly attractive public services–tax mix. The demand for that town's services–tax mix might well cause property values there to rise as people try to move in. A final equilibrium cannot be achieved until the relative attractiveness of the town's services–tax mix is fully capitalized into the value of the town's property.

The Pauly Model of the Housing Market

In general, the housing market has an important impact on both the nature of the equilibrium and whether an equilibrium even exists. A model developed by Mark Pauly is instructive for exploring the various possibilities when the jurisdictions are fixed (Pauly, 1976). It lies at the opposite end of the spectrum from McGuire's frontier model in assuming a fixed number of communities along with a fixed number of housing sites.

Let

X = a composite commodity whose price equals 1.
G = a bundle of public services, exclusive of taxes.
g = the unit price of (tax for) a public service, assumed constant across all jurisdictions.
R = the rental value of a standardized vector of property and housing services.
Y = lump-sum consumer income, assumed fixed for each individual.

Consumer utility is defined over X and G. Thus, consumers solve the following "as if" maximization problem, that is, as if they could choose the value of G:

$$\max_{(X_h, G_h)} \; U^h(X_h, G_h)$$
$$\text{s.t. } Y^h = X_h + gG_h + R \quad h = 1, \ldots, H$$

Assume initially that R is equal across all jurisdictions. The as-if maximization determines each person's most preferred amount of G, which they will try to match as closely as possible with the set of Gs offered in the given communities.

Suppose the maximization generates a G_h^*. Person h, and all other consumers identical to h in terms of preferences and income, will want to form a jurisdiction providing exactly G_h^* of public services, replicating if necessary to avoid any increases in R. If they could do so, preferences for public service bundles would be met exactly by homogeneous subgroupings of the population, and there would be no capitalization. This is the situation envisioned by the McGuire frontier model. Furthermore, g is essentially a head tax so that the subgroups would generate a pareto-optimal equilibrium.

Suppose, however, that there are a fixed number of localities, $\ell = 1,\ldots,L$ with the following characteristics:

1. Each locality offers a particular level of public services represented by the vector $\vec{\mathbf{G}} = (G_1, \ldots, G_\ell, \ldots, G_L)$ in ascending order of G_ℓ;
2. Each town has a fixed number of properties, represented by the vector $\vec{\mathbf{H}} = (H_1, \ldots, H_\ell, \ldots, H_L)$, such that $\sum_{\ell=1}^{L} H_\ell$ equals the entire population of individuals seeking a location; and
3. Rental values are specific to each locality, represented by the vector $\vec{\mathbf{R}} = (R_1, \ldots, R_\ell, \ldots, R_L)$.

In this case, it is possible that no individuals will find their preferred G_h^*, given the vectors of rental values and available public service bundles. All one can say is that individual h will locate in town l_1 if

$$V^h\left(G_{\ell_1}, R_{\ell_1}\right) > V^h\left(G_\ell, R_\ell\right), \quad \text{for } \ell \neq \ell_1 \tag{27.21}$$

where $V^h(\,)$ is the indirect utility function of person h. G and R are parameters from the individuals' point of view. Together, they determine X_h, given Y^h, from the budget constraint.

Let $\eta(G_\ell)$ = the number of people who choose to locate in locality ℓ, $\ell = 1,\ldots,L$. Equilibrium requires that

$$\eta(G_\ell) = H_\ell \quad \ell = 1, \ldots, L \tag{27.22}$$

Everyone has to live somewhere.

There will be no capitalization of public service bundles in equilibrium only if $R_\ell = \vec{\mathbf{R}}$, for $\ell = 1,\ldots,N$, holds as well. Return to the initial situation in which rental values are equal across all localities. Rental values can remain equal only if the search criterion (Eqn (27.21)) over all $h = 1,\ldots,H$ produces an exact matching of desired locations with the vector of locations available across all communities. Needless to say, a perfect matching without capitalization is unlikely. If it does not obtain, then the vector of rental values must change.

For example, suppose there existed a perfect matching that was upset by a sudden decline in G_1, the public services offered in the first locality. Some consumers in town 1, those who were closest to indifference between town 1 and 2, now prefer town 2 at the existing rental values. Their attempt to move to town 2 may drive up rental values there. But if rental values in town 2 begin to rise, some people in town 2, those closest to indifference between town 2 and town 3 at the initial equal rental values, now prefer town 3.

Rental values in town 3 may begin to rise, and so on. The rental values in all towns may change. Another possibility is that R_1 decreases as G_1 increases and everyone stays put. In any event, equilibrium can only be restored if Eqn (27.22) is reestablished for all l, and it will be an equilibrium with capitalization of public service bundles.

The equilibrium may never be restored, however. Pauly offers the following scenario as an intuitive counterexample relating to local educational services financed by local property taxes. Suppose there are two classes of otherwise identical families: small families with two or fewer children and large families with more than two children. Small families naturally want to live in towns with other small families; otherwise, the small families would be subsidizing the education of the large families for any given level of educational expenditures. Thus, if a given community consists of, say, an equal mix of large and small families, the small families will search for communities with a higher percentage of small families. But large families also prefer communities with a higher proportion of small families because of the resulting educational subsidies. Hence, large families follow the small families in their search for communities with a higher percentage of small families. As rental values adjust, small families move once again, only to be followed by the large families, and so on.

The system may reach an equilibrium if rental values in mixed communities exactly capitalize the pattern of subsidies, which are absent in the homogeneous communities. That is, the rental values of small homes in mixed communities would have to be less than the rental values of small homes in a homogeneous community of small families (given equal education expenditures) to offset the subsidy paid by small families in the mixed communities. Conversely, the rental values of large homes in mixed communities would exceed their rental values in homogenous communities of large families.[5] But even if such capitalization occurred, there is no guarantee that Eqn (27.22) can be satisfied as required for a general equilibrium.

5. Bruce Hamilton provides a similar example for high-income and low-income people. In his model, which uses a property tax, the low-income properties in mixed communities command a premium relative to their value in homogeneous communities, since their share of taxes declines in the mixed community for a given level of public services. The opposite holds for higher income properties in mixed communities. His model generates an equilibrium because properties can be expanded or contracted in each town, but it is not an efficient equilibrium. Because land values rise for low-income properties in the mixed communities, suppliers have an incentive to oversupply low-income housing in these communities. Consequently, low-income housing prices in mixed communities will no longer reflect the true value of the subsidies provided by the property taxes collected on the high-income properties. See Hamilton (1976).

The Hohaus–Konrad–Thum Model of Housing Market Distortion

The housing market is both beneficial and harmful in models of federalism. Its beneficial function is the one described above, that it helps to bring the sorting of people across fixed communities to an equilibrium. At the same time, however, the housing market introduces two potential sources of inefficiency into the economy that preclude the achievement of a first-best social optimum. One is that it gives local officials the option of levying a property tax, which is the easiest tax to collect at the local level and therefore the one that localities use in the United States and elsewhere. Unfortunately, a property tax is not a lump-sum tax. It introduces a standard second-best tax distortion by increasing the relative price of housing services. The second, more subtle, distortion is that it can prevent localities from providing the level of public goods that maximizes social welfare. We will consider the second distortion here because it is the one inherent in the sorting process.

Bolko Hohaus, Kai Konrad, and Marcel Thum analyzed this sorting distortion with a simple model patterned after Hotelling's model of optimal product differentiation (Hohaus et al., 1994). The model has the following elements:

1. *Fixed communities and housing sites.* Hohaus, Konrad, and Thum posit two equal-sized communities, L and H, with just enough housing sites to accommodate the entire population. The entire population is defined as a continuum indexed from (0, 1), and half the population must choose to live in each community.
2. *The public sector and political system.* Each community provides a Samuelsonian nonexclusive public good, X, excludable to members outside the community. The preferences for X are ordered along the same (0, 1) continuum as the population. That is, person j prefers X_j. The political system that determines the amount of X in each community is a direct democracy, with the median voter decisive. When voting for the public good in each community, the median voter takes as given the amount of the public good in the other community.
   The communities levy equal lump-sum head taxes to pay for the public good, whose total cost is C in each community. There is no tax distortion in the model.
3. *Social welfare.* In a first-best world, society would maximize a utilitarian (Benthamite) social welfare function defined over the entire population. This becomes the welfare standard against which the actual equilibrium is compared.

Preferences. The individuals have identical utility functions that depend on how closely the public good provided in their community matches their most preferred amount of the good. Letting X_H and X_L be the amounts of

the public good in communities H and L, the utility functions of individual i in H and individual j in L are

$$U(X_H, X_i) = \alpha - \beta(X_H - X_i)^2 \tag{27.23}$$

and

$$U(X_L, X_J) = \alpha - \beta(X_L - X_J)^2 \tag{27.24}$$

Hohaus, Konrad, and Thum assume that the people have distributed themselves across the community such that those with the lowest preferences for X are in L, (0, 0.5), and those with the highest preferences are in H, (0.5, 1), hence the use of L and H to designate the low- and high-X communities. As we will see below, this distribution is the one that brings the provision of X in the actual equilibrium closest to the provision that maximizes social welfare. Also, with this distribution of the people, the head taxes are C/0.5 in each community, since

$$\int_0^{.5} (C/0.5)\mathrm{d}X = C = \int_{0.5}^{1} (C/0.5)\mathrm{d}X \tag{27.25}$$

The Housing Market Equilibrium

In deciding which community to choose, people compare the housing prices and the amounts of the public good in each community. The housing market equilibrium must be such that the person on the margin is just indifferent between the two towns. Given the ordering of X, this is the person whose preferred amount of X is 0.5. Therefore, in equilibrium, the housing prices in the two communities, P_L and P_H, must satisfy

$$P_L + \beta(X_L - 0.5)^2 = P_H + \beta(X_H - 0.5)^2 \tag{27.26}$$

or

$$P_L - P_H = \beta(X_H - 0.5)^2 - \beta(X_L - 0.5)^2 \tag{27.27}$$

The Median Voter

Given the assumed distribution of the population, the median voters in the two towns would prefer $X_L = 0.25$ and $X_H = 0.75$, absent any consideration of housing prices. But the natural assumption is that they do care about housing prices. In particular, they care about the difference in the housing prices in the two communities should they ever decide to move to the other community. They also understand that the amount of X they choose affects the difference in housing prices. Therefore, they choose their optimal amount of X upon considering both their preferred amount of X and the difference in housing prices.

Consider the median voter in L. His or her goal is to maximize

$$\begin{aligned} U(X_H, X_L) = {} & P_L(X_H, X_L) - P_H(X_H, X_L) \\ & + \alpha - \beta(X_L - 0.25)^2 - C/0.5 \end{aligned} \tag{27.28}$$

But the housing prices are given by Eqn (27.27). Substituting Eqn (27.27) into Eqn (27.28) yields

$$\begin{aligned} U(X_H, X_L) = {} & \beta(X_H - 0.5)^2 - \beta(X_L - 0.5)^2 \\ & + \alpha - \beta(X_L - 0.25)^2 - C/0.5 \end{aligned} \tag{27.29}$$

Taking X_H as given, the first-order conditions with respect to X_L are

$$\partial U/\partial X_L = -2\beta(X_L - 0.5) - 2\beta(X_L - 0.25) = 0 \tag{27.30}$$

$$X_L = 0.375 \tag{27.31}$$

Similar analysis of the median voter's decision in H yields $X_H = 0.625$. Notice from Eqn (27.27) that with X_L and X_H both 0.125 removed from 0.5, there is no difference in housing prices in the two communities in the voting equilibrium.

The Social Welfare Optimum

The social welfare optimum under a utilitarian social welfare function would be $X_L = 0.25$ and $X_H = 0.75$, the median voters' preferred amounts of X. With the optimal Xs each 0.25 from 0.5, they would also imply no difference in housing prices from Eqn (27.27). This is the result that Tiebout had in mind: People with the closest preferences for the local public good would live together and provide the best possible match of the public good to their preferences. Instead, the actual equilibrium produces too much conformity in X relative to the social welfare optimum. The distortion arises because the person whose preferences determine the housing market equilibrium differs from the median voter in each community.

Note, also, that the assumed distribution of people across the two communities produces the largest difference between X_L and X_H. Suppose, instead, that people were uniformly distributed across the communities, the opposite extreme from the assumed distribution. Then the preferred X by the median voters in both communities would be 0.5, which, from Eqn (27.30), would be the amount of X provided in each community. This total conformity is as far as possible from the social welfare optimum of $X_L = 0.25$ and $X_H = 0.75$.

Sophisticated Voters

The Hohaus–Konrad–Thum model is convenient for demonstrating a point made earlier in the chapter, that more sophisticated voters can often improve the outcome of the

sorting mechanism. Consider again the median voter in L. Suppose this voter understands the structure of preferences well enough to realize that the voting equilibrium has to be symmetric around 0.5, so that $(X_H - 0.5) = (0.5 - X_L)$, or $X_H + X_L = 1$. This is hardly a great leap in sophistication, since Eqn (27.29) assumes that the voter understands how equilibrium is determined in the housing market. Substituting for X_H in Eqn (27.29) and maximizing yields $X_L = 0.25$. A similar understanding by the median voter in H yields $X_H = 0.75$. Hence, replacing the Nash assumptions with this additional degree of sophistication generates the social welfare optimum.

Empirical Estimates of Public Services Capitalization

Economists use hedonic price estimation to determine whether local public services and taxes are capitalized into housing prices as the Tiebout sorting theory suggests they should be. The results of these studies are mixed, and the technique fell out of favor as economists began to realize that no clear pattern of capitalization was ever likely to emerge.

With imperfect Tiebout sorting, there is no reason to expect any one pattern of coefficients on the public sector variables, either expenditures or taxes. They could be positive, negative, or zero, depending on people's preferences for expenditure–tax bundles relative to the bundles actually provided. If, for example, most communities in a given area are providing low-service bundles, whereas most people prefer high-service bundles, one would expect to find a positive correlation between property values and public services. If the situation were reversed, the regression coefficient would be negative. Worse yet, suppose most towns are offering either high or low levels of public services, whereas most people prefer a medium level of service. If the distribution of public services were symmetric across communities, a regression of property values on public services would yield a zero coefficient. Yet theory would suggest that rental values in the medium service communities would capitalize the excess demand for these services. Thus, even if capitalization is occurring, regression analysis may fail to discover it.[6]

Finally, if Tiebout sorting were perfect such that all households lived in communities with exactly the public service bundle they desired, then the hedonic price estimates on the public sector variables would again yield zero coefficients.[7] In short, economists realized that testing for capitalization with the hedonic price techniques was unlikely to generate useful information.

Tiebout Sorting from an Historical Perspective

Tiebout's conjecture that people will choose localities in part on their public service–tax mix generates two testable hypotheses about the nature of localities as people move over time. The first is that localities will become more homogeneous with respect to people as people with like tastes will want to live together. The second, which follows from the first, is that the public service–tax mix will become more heterogeneous across localities. For example, young couples with children will want better schools and be willing to pay higher taxes for them; older adults whose children have grown will want more emphasis on safety and cultural events, which can be provided more cheaply than schools. So young couples live together, the older adults live together, and the public service–tax mix in their separate communities will differ. Moreover, Tiebout sorting should work even better if moving costs and the costs of remaining in contact decrease over time.

Paul Rhode and Koleman Strumpf tested these hypotheses on three huge data sets of demographic and public sector variables in the United States: (1) All the municipalities (i.e., cities and towns) in a random sample of 10% of all US counties from 1870 to 1990, (2) All counties in the United States from 1850 to 1990, and (3) Ninety-two municipalities in the greater Boston area from 1870 to 1990. They chose a number of proxies for tastes such as age, race, religious affiliation, voting shares by party in Presidential elections, and, from 1970 to 1990 when they became available, per capita income, education levels, and rates of home ownership. The public sector variables included total taxes and spending, spending on schools and the associated school taxes, spending on protection services such as fire and police, and other spending categories.

They noted that the costs of moving and communicating decreased dramatically over this time period. Nonetheless, they could find no evidence to support either hypothesis. Indeed, quite the opposite was true. No matter what combination of taste proxies they chose, municipalities and counties became more heterogeneous with respect to tastes over the sample period. Likewise, the public service/tax mix became more homogeneous across the municipalities, probably because they became more heterogeneous with respect to tastes. They conclude that Tiebout sorting has not been an important phenomenon in the United States.[8]

6. This particular example is due to Pauly, although Bruce Hamilton has made essentially the same point. See Pauly (1976), Hamilton (1976).
7. This point was made, and tested, by Matthew Edel and Elliot Sclar in Edel and Sclar (1974).
8. Rhode and Strumpf (2003). The one exception in their results is that the municipalities in the greater Boston area became more stratified by race, but there is ample evidence that this was the result of discrimination, not Tiebout sorting.

ANYTHING IS POSSIBLE

Theoretical models of fiscal federalism have shown that the sorting of people can lead to almost any outcome. We conclude this section with a simple model by Stiglitz that can produce a wide range of outcomes, from an efficient equilibrium to an inefficient equilibrium, multiple equilibria, or an equilibrium that can be improved upon by grants-in-aid from high-income to low-income communities.[9]

The Stiglitz Model

The Stiglitz' model is a variation of the McGuire model. Stiglitz posits a fixed number of communities, but with sufficient undeveloped land available in each community that expansion or contraction of the town has no effect on land prices. Thus, he does not include a housing market. The main difference from the McGuire model is that income is generated by production in the communities, with each person supplying one unit of labor. There are no other factors of production. The output from production can take the form of a private good, X, and a Samuelsonian nonexclusive good G. A second difference from the McGuire model is that G is not subject to congestion. Its services are equally available to everyone in the town. Finally, mobility from community to community is costless, so that horizontal equity—equal treatment of equals—is the sorting equilibrium condition. Since the people are assumed to be identical, everyone must have the same utility in equilibrium.

Production

Let

$$Y = f(N),\ f' > 0 \text{ and } f'' < 0 \tag{27.32}$$

define the total income or output generated by the N people in the community according to the production function $f(N)$. Also,

$$f(N) = NX + G \tag{27.33}$$

where X is the amount of the private good received by each person.

Preferences

The identical individuals have utility functions defined over X and G:

$$U = U(X, G) \tag{27.34}$$

Also, Eqn (27.33) implies that the budget constraint for each individual is

$$f(N)/N = X + G/N \tag{27.35}$$

The first task is to describe the optimal levels of G and N.

The Optimal G

To determine the optimal amount of G for a given N, the individuals solve the following problem:

$$\max_{(X,G)} U(X, G)$$
$$\text{s.t. } f(N)/N = X + G/N$$

with the corresponding Lagrangian:

$$\max_{(X,G)} U(X, G) + \lambda(f(N)/N - X - G/N)$$

The first-order conditions are

$$X : U_X - \lambda = 0 \tag{27.36}$$

$$G : U_G - \lambda(1/N) = 0 \tag{27.37}$$

Dividing Eqn (27.37) by Eqn. (27.36) and rearranging terms yield

$$N(U_G/U_X) = 1 \tag{27.38}$$

the standard Σ MRS = MRT condition for nonexclusive goods.[10]

The Optimal N

In this model, N appears only in the budget constraint of the consumer problem. Therefore, the N that maximizes utility is the N that maximizes X for a given G (or vice versa). Write the budget constraint as

$$X = (f(N) - G)/N \tag{27.39}$$

The maximum X for a given G is given by

$$\partial X/\partial N = \left(Nf' - f(N) + G\right)/N^2 = 0 \tag{27.40}$$

or

$$Nf' - f(N) + G = 0 \tag{27.41}$$

or

$$f' = (f(N) - G)/N = X \tag{27.42}$$

Equation (27.42) says that the community should expand to the point at which the marginal product of the last person just equals his consumption. This is the equivalent of McGuire's idea that new entrants have to pay their

9. Stiglitz (1977). Reproduced in abridged form in Aktinson and Stiglitz (1980).

10. The MRT is defined in terms of NX along the aggregate production-possibilities frontier, Eqn (27.33).

marginal costs of entering a community so that the existing residents are willing to accept them. In the Stiglitz model, there is no reduction in G or in the private good available to anyone else if the last person in the community produces just enough to cover his own consumption. This is a common result in the literature.

The Henry George Theorem

Equation (27.41) points to another common result known as the Henry George theorem. Rewrite Eqn (27.41) as

$$G = f(N) - Nf' \tag{27.43}$$

With labor paid its marginal product, the term Nf' on the right-hand side (RHS) of Eqn (27.43) is the total wage bill, and the entire RHS is the economic profit from production. Therefore, Eqn (27.43) implies that the public good should be paid for by a 100% tax on economic profits, a nondistorting lump-sum tax. This result is called the Henry George theorem after Henry George, a New York City politician in the late 1800s who led a "single-tax movement" to finance local government spending with a tax on land. He argued that a tax on land, or equivalently on annual land rents, would be nondistorting since the supply of land is fixed. Taxing pure economic profits is equivalent to taxing land rents in the sense that they are both nondistorting. Many models of federalism with local production generate this result.

Community Formation: Varying *N* and *G*

As new people enter an existing community and join in production, the amount of X and G varies. A central feature of the model is that the opportunity locus of X and G available to each individual is concave as N varies. Figure 27.2 illustrates.

The individual budget line for a given N is $X = f(N)/N - (1/N)G$, with slope $= -1/N$. As N increases, the maximum possible X, $f(N)/N$, decreases with $f'' < 0$ and the maximum amount of G, $f(N)$, increases. In addition, the maximum utility available to the individual along each budget line first increases and then decreases. At low N, utility is low because G is low. At high N, utility is low again because X is low. Congestion in this model occurs in terms of X, not G.

Possible Equilibrium Outcomes

Define $V(N)$ as the maximum utility attainable at each N. Stiglitz illustrates different equilibrium outcomes depending on the precise shape of $V(N)$, with each equilibrium satisfying horizontal equity across communities.

Efficient Equilibrium

Suppose $V(N)$ is symmetric as in Fig. 27.3, reaching its peak at N^*. The individuals will sort themselves in communities of size N^*. If the total population is an even multiple of N^*, then the sorting equilibrium is a pareto optimum. In order to compare the efficient case with the other possibilities, assume that the total population, $\overline{N}$, is twice N^*, with two communities in equilibrium.

Inefficient Equilibrium

A variety of problems can arise if $V(N)$ is asymmetric. Suppose that $V(N)$ is as pictured in Fig. 27.4, and consider how the people will sort themselves into two communities. N_1 goes to the right and N_2 to the left, with $N_1 + N_2 = \overline{N}$. $V(N)$ can no longer be maximized in both communities. The best outcome is point B with half the population in each community, but society may not get there. Suppose an initial sort occurs to the right of B, at N_1^1 and N_2^1. Utility is higher in community 1, so people will leave community 2

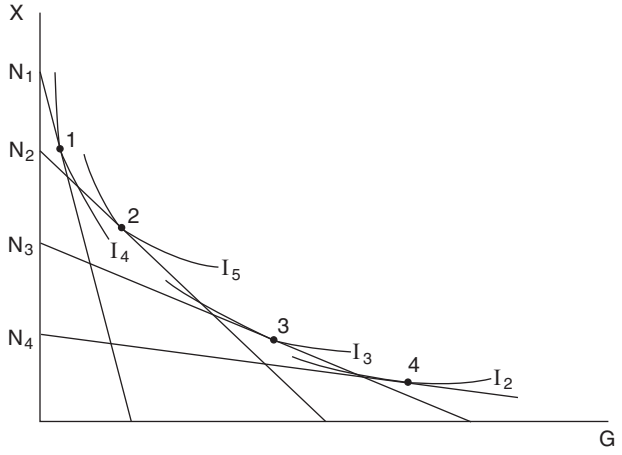

FIGURE 27.2

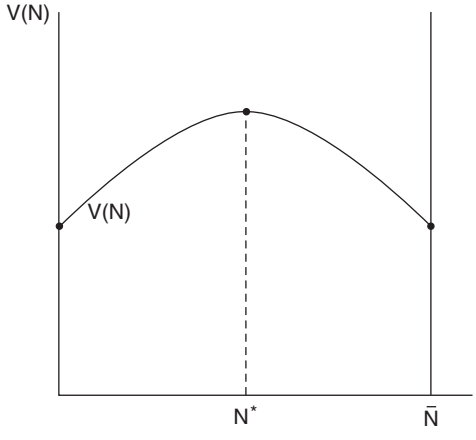

FIGURE 27.3

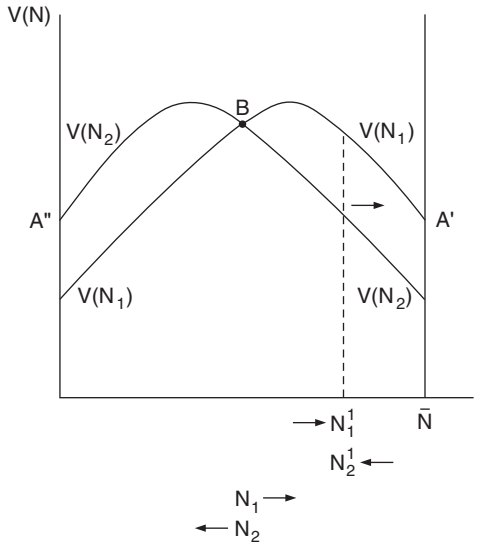

FIGURE 27.4

for community 1. The movement continues until everyone resides in community 1, with utility of A′. Similarly, if the initial sort is to the left of B, then everyone will move to community 2, with utility A″. A′ and A″ are both pareto inferior to B.

Efficient Equilibrium, But Unstable

Figure 27.5 pictures a variation of the previous case. Here, the best equilibrium is the pareto optimum C, with $N_1 = N_2 = N^*$, but it is not a stable equilibrium. The slightest movement away from C will lead everyone to reside in one community, with utilities A′ or A″. Instead of an optimal federalism, society is likely to get inefficient national provision of the public good.

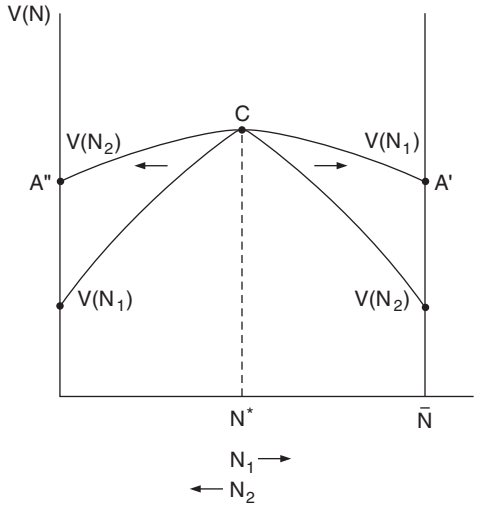

FIGURE 27.5

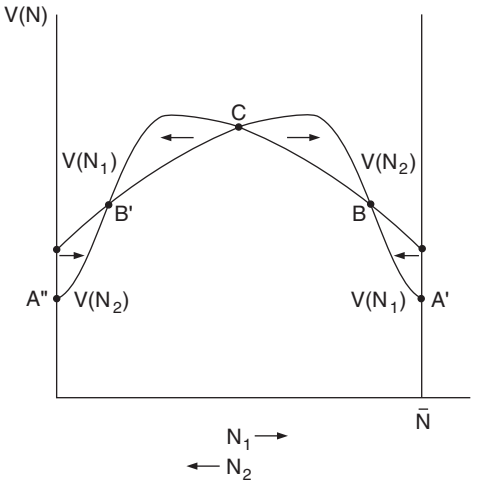

FIGURE 27.6

Multiple Equilibria

Figure 27.6 illustrates the case of multiple equilibria, each with two communities. C is the best outcome, but B and B′ are the only stable equilibria. People move to B′ from anywhere to the left of C and to B from anywhere to the right of C.

Grants-in-Aid

The final example suggests a possible role for grants-in-aid among communities on efficiency grounds. Suppose the economy consists of two types of communities with different production technologies. $V(N)$ for each community type is pictured as the two solid lines in Fig. 27.7. $V(N)$ is symmetric, with a value of zero at $N = 0$ and at $\widehat{N}$. There are no longer just two communities.

Everyone will live in a high-productivity town if there are enough such communities to accommodate everyone and give them utility greater than A, the maximum utility

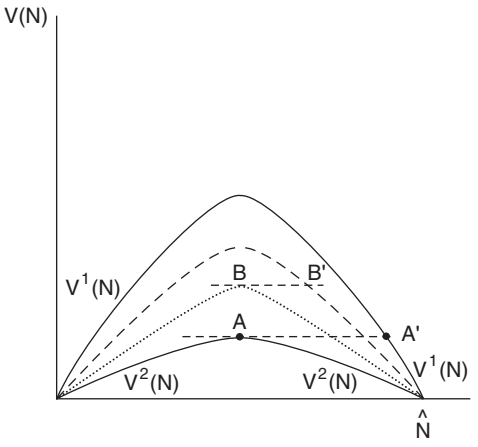

FIGURE 27.7

attainable in the low-productivity communities. If there are not enough high-productivity communities, however, a horizontal-equity equilibrium with utility of A = A′ in the two communities will obtain. The more productive community is the larger one. Society can do better if the more productive communities subsidize the less productive communities with a grant-in-aid, shifting the $V(N)$ lines to the dotted lines in the figure. A new horizontal-equity equilibrium obtains with utility of B = B′, greater than A = A′. The loss in income in the high-productivity communities from the grant-in-aid is more than compensated for by the movement of people to the low-productivity communities. This is another example of how a more sophisticated citizenry can improve the efficiency of the sorting process.

The ability of a simple model to generate such a variety of outcomes is unsettling and fairly devastating to Tiebout's optimistic conjecture about the potential advantages of a federalist system of government.[11,12] That said, Stiglitz's model produces three results that are common in the federalism literature, including:

1. The Σ MRS = MRT rule for allocating the local public good.
2. The last entrant into the community pays the marginal cost of his or her entry at the optimum.
3. The Henry George theorem, which calls for 100% taxation of economic profits or land rents to pay for the public good when the model includes local production.

MOBILITY AND REDISTRIBUTION

We noted in Chapter 26 that the prevailing model of fiscal federalism calls for redistribution by the national government, in part because of the so-called competition problem. Mobility is seen to undermine the redistributional efforts of the lower level ("local") governments. A fairly large literature has evolved that explores in a positive vein the implications of local redistributions (LRs) versus national redistribution when mobility is possible. The overall message from these studies is that mobility restricts but does not entirely destroy the possibilities for redistributions by lower level governments. These governments can likely engage in a considerable amount of redistribution even if mobility were costless. Furthermore, national redistribution is not necessarily preferred to LRs. We conclude this chapter with a discussion of three studies that highlight the main issues involved in comparing national versus LR.

The Brown–Oates Model

Charles Brown and Wallace Oates published one of the first studies comparing local with national redistribution in 1987, a study that is still widely cited (Brown and Oates, 1987). They adapted the pareto-optimal redistribution model of altruism to a federalist setting. In their model, each locality i consists of N_i nonpoor and P_i poor individuals. The nonpoor are altruistic to the poor—their utility depends upon their own income and the income of the poor. But the altruism of the nonpoor extends only to the poor within their locality, that is, redistribution is a local public good. This assumption generates a motivation for LR. Transfers are given equally to the poor within each locality and are financed by equal lump-sum taxes on the nonpoor. Thus, under LR the cost to each of the nonpoor of a dollar of transfer to each poor person in locality i is the ratio of the poor to the nonpoor, (P_i/N_i).

The poor care only about their own income, and they are the only mobile citizens in the model. Mobility is costly, consisting of one component, α, common to all the poor and another component that is specific to each poor person. Define C_k^i as the specific migration cost of person k in locality i. Poor person k will migrate from locality i to locality j if $T^j - T^i > C_k^i + \alpha$.

To isolate the effect of mobility on the possibilities for LRs, Brown and Oates assume that all nonpoor people have identical tastes and income and that all poor people also have identical tastes and income. Under these assumptions, differences in the amounts of transfer across localities under LR depend only on the ratio (P_i/N_i) in each locality, the price of the transfer to the nonpoor. The transfer is determined by a majority vote, but the nonpoor are always assumed to be in the majority ($N_i > P_i$, all i). Finally, the nonpoor in each locality select their desired transfer to the poor under the Nash assumption that transfers in all the other localities remain at their current levels.

Local redistribution under these assumptions is compared with national or centralized redistribution (CR). Under CR, all the nonpoor are assessed equal head taxes to pay for transfers to all the poor. Thus, the price of a dollar of transfer to the poor is the same for all the nonpoor, equal to the ratio of the total poor population to the total nonpoor population, $\left(\frac{\sum_i P_i}{\sum_i N_i}\right)$. The poor everywhere receive the same transfer and there is no incentive for mobility.

The question Brown and Oates ask is whether the average level of transfers to the poor is lower under LR than under CR in this model, and the answer is "not necessarily." Consider the case of just two localities. Under LR with the Nash assumption, mobility leads to a general

11. One ray of hope is that adding congestion of the public good to the Stiglitz model increases the likelihood of reaching a pareto optimum with two or more communities.

12. Those interested in pursuing the sorting-of-people literature further might consult the following articles: Bewley (1981) (one of the first articles to formally question Tiebout's conjecture), Epple et al. (1993), Henderson (1985) (shows that profit maximization by local developers can be pareto optimal).

incentive for the nonpoor to reduce the amount of transfer. An other-things-equal increase in locality i's transfer to the poor induces immigration of some poor from the other locality, j, which increases (P_i/N_i), the price of the transfer. The locality with the higher transfer, say i, experiences immigration of new poor, so the price actually does rise and the transfer falls. In locality j, however, the price of transferring (P_j/N_j) falls, which overcomes the general incentive not to raise the transfer and leads to an increase in its transfer. Therefore, with the transfer rising in one locality and falling in the other, the average level of transfers on LR could exceed the level of the transfers under CR.

Simulation Results

Brown and Oates perform a simulation with two localities to test whether LR is likely to reduce the average level of transfers. The nonpoor have a CES utility function defined over their own income and the income of the poor in their locality. The endowment income of the poor is set at 1/4 of the endowment income of the nonpoor. Initially, 60% of the poor reside in one community and 40% in the other. The specific cost component of mobility, C_k^i, is assumed to be normally distributed with a mean and variance such as to distribute the population initially in the 60/40% ratio.

The simulations produce two expected results:

1. LR is likely to generate a lower average level of transfer than CR (this result requires that the elasticity of substitution of the nonpoor between their own income and the income of the poor be less than 1, which is the expected range).
2. Increased mobility leads to lower average levels of transfer under LR (mobility is increased by lowering the general mobility cost component a from ∞ (no mobility) to 0).

A more surprising result is that the average transfer is higher under CR than under LR even with no mobility. This occurs because the price under CR, $(P_1 + P_2)/(N_1 + N_2)$, turns out to be lower than the average of the prices in the two localities, (P_1/N_1) and (P_2/N_2), in their simulation model (again, assuming that the elasticity of substitution in the CES utility function is <1).

Brown and Oates conclude that their simulations indicate a preference for CR over LR even when the motivation for redistribution is local. In addition to higher average transfers, CR avoids differential treatment of the poor across localities. Under LR, the amount of transfer differs considerably in the two localities in all their simulations. This differential treatment is unsettling since it depends so heavily on the initial distribution of the poor across the localities. One common objection to the US public assistance programs is that the benefits that the poor receive depend on which state they happen to live in.

At the same time, however, the simulations suggest that LR can lead to substantial redistributions, even when the costs of mobility are as low as possible ($\alpha = 0$). The differences in the average levels of transfer under the two regimes are never huge. The case for national redistribution is suggestive but by no means conclusive.

Uncertain Incomes

In 1998, Kangoh Lee published a modification of the Brown–Oates model that allows for the possibility of uncertain incomes (Lee, 1998). The uncertainty comes from favorable or adverse shocks to local production, which may be idiosyncratic to localities or national in scope, affecting production everywhere equally. Uncertain local income generates a further presumption in favor of national over local redistribution because national redistribution provides insurance against idiosyncratic shocks.

In Lee's model, each locality i consists of one rich person and N_i poor people, with the rich person being altruistic toward the poor within his or her own locality, as in the Brown–Oates model. Unlike the Brown–Oates model, however, the altruism of the rich toward the poor can vary across localities.

Production occurs in each locality according to the production function:

$$Y_i = \theta_i f(N_i) \tag{27.44}$$

θ_i is a random productivity shock that takes on the values $\underline{\theta}$ with probability p, and $\overline{\theta}$ with probability $(1 - p)$. The poor workers receive wages equal to their marginal products, $\theta_i f'(N_i)$. The rich person owns the production process (or an unnamed fixed factor) and receives the profits $g(N_i) = \theta_i(f(N_i) - f'(N_i))$. The output from production, Y_i, is a composite commodity whose price is 1.

Given that the preferences of the rich toward the poor can differ, Lee's benchmark presumption is that LR is preferred to CR in a world of no mobility and certain incomes. Pareto-optimal redistribution requires varying transfers across localities, which is not possible under CR, by assumption. As in the Brown–Oates model, CR implies equal transfers to all the poor.

Uncertainty greatly complicates the analysis. To capture the intuition of how the model works, consider the case of just two localities with equal numbers of workers, n, in each locality. Lee assumes that the local or centralized transfers to the poor are set before the production shocks are realized. Thus, the objective functions under LR and CR are

$$\text{LR}: \max E\left[U_i^R(Y_i, y_i)\right] \text{ in each locality}, \quad i = 1, 2$$
$$\text{CR}: \max E\left[U_1^R(Y_1, y_1)\right] + E\left[U_2^R(Y_2, y_2)\right]$$

where Y_i is the income of the rich person and y_i the income of each poor person in locality i. One important difference from the Brown–Oates model is that the tax to pay for the

transfers is a tax on the income of the rich, a more realistic assumption.

No Mobility

Begin with the case of no mobility to focus on the insurance advantage of CR. Further, assume that the preferences of the rich are identical, so that the poor receive the same transfer under either LR or CR. (Recall that the transfers are set before the production shocks occur.) There is no efficiency advantage to LR absent uncertainty. Suppose, first, that the shocks are identical in the two localities, either adverse ($\underline{\theta}$) or favorable ($\overline{\theta}$) in both (a national shock). CR has no insurance advantage in this case, so society is indifferent between LR and CR under the given assumptions.

Suppose, instead, that the shocks are idiosyncratic, either $(\underline{\theta}_1, \overline{\theta}_2)$ or $(\underline{\theta}_1, \overline{\theta}_2)$. These outcomes happen with equal probability $p(1-p)$. Since the transfers are set before the shock occurs, they are equal under LR or CR. Let the transfer be T per poor person.

Under LR, the rich in each locality pay taxes equal to nT regardless of the shock to the community. The total taxes paid by the rich are $2nT$. Under CR, in contrast, the income tax payments to support the transfers vary according to the shocks:

$$\text{Adverse} - \text{shock locality} : \text{Tax} = 2nT\left(\frac{\underline{\theta}}{\underline{\theta}+\overline{\theta}}\right) < nT$$

$$\text{Favorable} - \text{shock locality} : \text{Tax} = 2nT\left(\frac{\overline{\theta}}{\underline{\theta}+\overline{\theta}}\right) > nT$$

Since these two outcomes occur with equal probability, the expected tax of each rich person is nT, equal to the actual tax under LR. But by redistributing the tax burden from the adverse- to the favorable-shock locality, CR reduces the variation in the rich people's after-tax incomes relative to LR. Assuming the rich are risk averse, this mean-preserving contraction of their uncertain incomes increases their utility. CR is preferred to LR; it is the pareto-optimal solution.[13]

The insurance advantage of CR may not be enough to overcome the inherent efficiency advantage of LR if the altruistic preferences of the rich differ. In general, however, with no mobility and uncertain incomes, the following is true: CR is more likely to be preferred to LR the more similar the preferences of the altruistic rich are and the more likely idiosyncratic the shocks are.

13. If the tax rates were set prior to the shock rather than the transfer payment, then the insurance advantage would be received by the poor. Under CR, they receive equal transfers. Under LR, the transfers would be lower in the adverse-shock locality and higher in the favorable-shock locality given the budget constraint.

Mobility

Lee considers the case of perfect mobility of the poor, which implies horizontal equity. In the context of his model, the income of the poor including the transfer must be equal no matter where they live.

Horizontal Equity Condition

$$\theta_i f'(N_i) + T_i = \text{all } i$$

Local Redistribution has disadvantages and advantages relative to CR under costless mobility. On the one hand, LR gives rise to two forms of inefficiency that are absent under CR: production inefficiency and Nash inefficiency.

Production Inefficiency

Consider two localities i and j that receive the same production shock. Suppose the rich person in i is more altruistic than the rich person in j and sets $T_i > T_j$. The poor workers will migrate from j to i until the horizontal-equity condition holds, which implies that $\theta_i f'(N_i) < \theta_j f'(N_j)$. But output is maximized by equalizing the marginal products across the two localities. The unequal degrees of altruism generate a production inefficiency.

Nash Inefficiency

Lee adopts the usual Nash assumption that each rich person sets the transfer under the assumption that the transfers in all other localities are being held constant. The rich understand that higher transfers lead to immigration of some poor and a higher tax burden, which reduces the incentive to provide as much transfer. But they miss the externality that their transfer decision imposes on the other rich as the poor leave the other localities. The externality has two dimensions. The other rich face lower tax burdens with fewer poor, yet the income generated in the other localities is also lower. Ignoring this externality is the so-called Nash inefficiency under LR.

On the other hand, LR has two advantages relative to CR: a redistributional efficiency advantage and a particular kind of insurance advantage.

Redistributional Efficiency

LR permits unequal transfers across localities, which is in itself utility enhancing relative to the single transfer under CR if the altruistic preferences of the rich vary.

Insurance Advantage

The mobility of the poor in response to idiosyncratic shocks performs an income insurance function for the economy by reallocating resources from the adverse-shock communities to the favorable-shock communities. This insurance

property of mobility greatly offsets the insurance advantage of CR, which operates through redistributing tax burdens among the rich in the presence of idiosyncratic shocks.

The net effect of the advantages and disadvantages of LR relative to CR is such that one regime is not necessarily preferred to the other. For example, Lee shows that CR may not be more efficient than LR even with identical preferences among the rich. Conversely, LR may not be more efficient than CR with varying preferences and identical (national) shocks because of the production and Nash inefficiencies that it gives rise to.

In general, Lee's analysis indicates that the case for CR versus LR with uncertain incomes turns on

1. The extent of heterogeneity in the altruistic preferences of the rich across localities.
2. The extent to which production shocks are national or idiosyncratic.
3. The extent of the mobility of the poor in response to differences in transfers across localities.

The Epple–Romer Model of Redistribution

The final model, by Dennis Epple and Thomas Romer, analyzes the possibilities for LR in a much richer environment than the two previous models (Epple and Romer, 1991). The Epple–Romer model is in the style of the Pauly model described earlier. They posit a set of households defined over a continuum of endowed income who must locate themselves within J local communities. Their utilities are a function of a numeraire composite commodity and housing. Although the number of communities is fixed, the supply of housing within each community is a variable, so that the number of people living in each community is endogenous. Mobility is costless.

The political process is a direct democracy with the median voter decisive. People in each community vote for a grant to be given equally to all residents, financed by a property (housing) tax. In deciding on the tax-transfer policy, voters are aware of its effect on housing prices, but they adopt the Nash assumption that their votes have no effect on the tax-transfer policies in the other communities. The other policies are taken as given. Notice that the redistributional motive in this model is entirely self-serving. The lower income residents can effect a redistribution from the higher income residents through the political process.

A model with all these features is highly complex. It requires that four conditions hold simultaneously for an equilibrium, three internal and one external. The internal equilibrium conditions are that, within each community, there must be

1. *Housing market equilibrium*—The demand for housing must equal the supply of housing. The consumers' demand for housing is a function of the price of housing gross of the property tax, and the supply of housing is a function of the price of housing net of the tax.
2. *Budgetary balance*—The sum of the grants must equal the revenues collected from the property tax.
3. *Voting equilibrium*—The tax-transfer combination is that preferred by the median voter, and it must be consistent with the other equilibrium conditions.

The external equilibrium condition is that no one wants to move to another community.

Epple and Romer make some realistic assumptions about consumers' preferences that ensure the existence of a full equilibrium. They consider two versions of the model. In the first version, all households are renters who pay rent to absentee landlords. In the second version, some or all households are homeowners.

All Renters

Begin with the all-renters model and consider the conditions on preferences that drive the results of the model. Utility is defined over a composite commodity (b) and housing (h). The budget constraint of a household living in community j is

$$y + g^j = p^j h + b \tag{27.45}$$

where g^j is the per-household grant in community j, and p^j is the gross-of-tax price of housing. Utility maximization leads to an indirect utility function defined over y, p^j, and g^j. Assuming housing is a normal good, the indifferences curves for g and p are as pictured in Fig. 27.8. Furthermore, at any given (g, p) combination, the indifference curves are flatter the higher the household's income, also as pictured. This condition on the marginal rates of substitution is crucial to the results of the model.

To see why, refer to Fig. 27.9. Suppose the household with income y is indifferent between the combinations (g^i, p^i) and (g^j, p^j). Then, any household with income greater than y would prefer (g^i, p^i) to (g^j, p^j), as illustrated

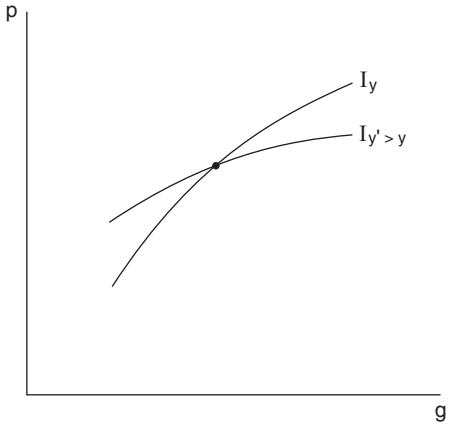

FIGURE 27.8

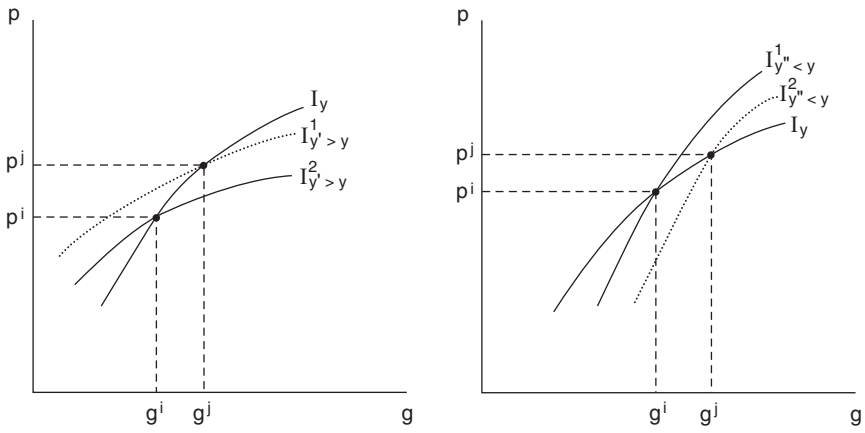

FIGURE 27.9

in the left-hand panel. Higher income households buy more housing and therefore require a bigger increase in g to compensate for the increase in p from p^i to p^j. Conversely, any household with income less than y would prefer (g^j, p^j) to (g^i, p^i), as illustrated in the right-hand panel. They require a smaller grant in compensation for the same increase in p.

Most households cannot achieve their most desired (g, p) combination with only a fixed number of communities to choose from. The following must be true, however, given the condition on the MRS: If two households with incomes y_1 and y_2, $y_2 > y_1$, most prefer community i with combination (g^i, p^i) from the J available choices, then all households with $y_1 < y < y_2$ also most prefer community i.

Finally, suppose that y_i is the highest income in community i and y_j is the highest income in community j. If $y_i > y_j$, then from the shape of the indifference curves, $g_i < g_j$ and $p_i < p_j$ in equilibrium. These considerations suggest the three principal results of the all-renter model:

1. The J communities stratify by income level.
2. The higher the grants and property taxes the lower the average income of the community. Lower income communities engage in more redistribution than higher income communities.
3. Even relatively high-income communities are likely to engage in some redistribution given that the preferences of the median voter are decisive. The Epple–Romer model has the property that the median voter over (g, p) within each community is the voter with the median income. Therefore, the median voter and all those with incomes below the median are likely to prefer some positive (g, p) even in very-high-income communities. Epple and Romer conclude that quite a lot of LR is possible even if mobility is costless.

Homeowners

Adding homeowners to the model dramatically reduces the incentive to redistribute at all income levels because of the effect of the property tax on the net-of-tax price of housing. An increase in the property tax lowers the net-of-tax price of housing, which hurts the absentee landlords in the all-renter model. The amount of land that the landlords own is fixed, as is the number of communities, so they cannot respond to the taxes. They end up bearing some of the burden of the redistribution. Homeowners, however, now bear this burden as a capital loss, and they can influence the voting outcome. Their budget constraint is

$$y + g + \left(p^{\text{net}} - p_0^{\text{net}}\right)h_0 = ph + b \tag{27.46}$$

where p and p^{net} are the gross- and net-of-tax housing prices, h_0 refers to the existing house, and p_0^{net} is the net-of-tax price paid originally for the house. The decline in p^{net} as taxes are raised to pay for the grant represents a capital loss, which homeowners take into consideration when voting on their preferred (g, p) combination. The capital loss flattens homeowners' indifference curves at every (g, p) combination relative to renters.

Simulation Results

Simulations of their model with three communities, roughly calibrated to the US economy, generate all these results. In one version, Epple and Romer constrain the highest income community not to redistribute. In the all-renter case, the middle-income community imposes a 28% tax on housing services to finance a grant of $1676 against a mean income of $21,560. It is not true that the competition problem constrains only the lowest income communities to redistribute. Even more striking is the reduction of redistribution to trivial amounts in the all-homeowners version; grants fall to a range between $68 and $133. Epple and Romer conclude that home ownership and not mobility may be the biggest hindrance to LRs in a federalist system.

How applicable these findings are for the United States or any other developed market economy is difficult to say. One caveat is that the propensity to vote is inversely related

to income, at least in the United States, so that the voter with the median preferences almost certainly does not have the median income in any community. Another is that LR is not compared with centralized distribution. Finally, the Epple–Romer model lacks any degree of altruism toward the poor. These last two points make it difficult to compare the Epple–Romer model with the other two models above, a comment that applies generally to the federalism literature. Models of federalism have so many dimensions to choose from that no consensus model of federalism has emerged or is likely to emerge. At best we are left with a number of suggestive results, with no way yet of achieving a satisfactory synthesis.[14]

REFERENCES

Aktinson, A., Stiglitz, J., 1980. Lectures on Public Economics. McGraw-Hill, New York (Chapter 17).

Bewley, T., May 1981. A critique of Tiebout's theory of local public expenditures. Econometrica 49 (3), 713–740.

Brown, C., Oates, W., April 1987. Assistance to the poor in a federal system. Journal of Public Economics 32 (3), 307–330.

Buchanan, J., February 1965. An economic theory of clubs. Economica.

Epple, D., Romer, T., August 1991. Mobility and redistribution. Journal of Political Economy 99 (4), 828–858.

Epple, D., Filimon, R., Romer, T., November 1993. Existence of voting and housing equilibrium in a system of communities with property taxes. Regional Science and Urban Economics 23 (5), 585–610.

Epple, D., Sieg, H., August 1999. Estimating equilibrium models of local jurisdictions. Journal of Political Economy 107 (No.4), 645–681.

Edel, M., Sclar, E., September/October 1974. Taxes, spending, and property values: supply adjustment in a Tiebout–Oates model. Journal of Political Economy 82 (5), 941–954.

Hamilton, B., June 1976. The effects of property taxes and local public spending on property values: a theoretical comment. Journal of Political Economy 84 (3), 647–650.

Hamilton, B., December 1976. Capitalization of intrajurisdictional differences in local tax prices. American Economic Review 66 (5), 743–753.

Henderson, J., April 1985. The tiebout model: bring back the entrepreneurs. Journal of Political Economy 93 (2), 248–264.

Hohaus, B., Konrad, K., Thum, M., 1994. Too much conformity? a hotelling model of local public goods supply. Economic Letters 44 (3), 295–299.

Lee, K., September 1998. Uncertain income and redistribution in a federal system. Journal of Public Economics 69 (3), 413–433.

McGuire, M., January/February 1974. Group segregation and optimal jurisdictions. Journal of Political Economy 82 (1), 112–132.

Pauly, M., October 1976. A model of local government expenditure and tax capitalization. Journal of Public Economics 6 (3), 231–242.

Rhode, P., Strumpf, K., December 2003. Assessing the importance of tiebout sorting: local heterogeneity from 1850 to 1990. American Economic Review 93 (5), 1648–1677.

Scotchmer, S., 1994. Public goods and the invisible hand. In: Quigley, J., Smolensky, E. (Eds.), Modern Public Finance. Harvard University Press, Cambridge, MA. Chapter 4.

Stigler, G., 1957. Tenable range of functions of local government. In: Federal Expenditure Policy for Economic Growth and Stability. Joint Economic Committee, Subcommittee on Fiscal Policy, Washington, D.C.

Stiglitz, J., 1977. The theory of local public goods. In: Feldstein, M., Inman, R. (Eds.), The Economics of Public Services: Proceedings of a Conference Held by the International Economic Association at Turin, Italy. Macmillan, New York.

Tiebout, C., October 1956. A pure theory of local expenditures. Journal of Political Economy 64 (5), 416–424.

14. Dennis Epple and Holger Sieg present an estimating procedure for testing the implications of these models that incorporates the equilibrium conditions in Epple and Sieg (1999).

Chapter 28

The Role of Grants-in-Aid in a Federalist System of Governments

Chapter Outline

OPTIMAL FEDERALISM AND GRANTS-IN-AID: NORMATIVE ANALYSIS

First-Best Policy Environment

Whether grants-in-aid have any role in an optimal, first-best federalist system of governments depends upon the underlying model used to establish the notion of a social welfare optimum. Recall that in the conventional model of optimal federalism redistributional policy is the sole responsibility of the national government, whereas allocational functions reside in the lowest level governments consistent with pareto optimality. Consequently, only the national government is concerned with social welfare optimization as traditionally defined. The lower level governments care only about efficiency.

Grants-in-aid are unnecessary in this model, as long as the policy environment is truly first best and a perfect correspondence of jurisdictions exists for all allocational problems. The national government satisfies its interpersonal equity conditions with lump-sum taxes and transfers among individuals (and firms, with decreasing cost production), exactly as in the single-government model of the public sector. Similarly, all governments, whether national or "local," interact only with the individual consumers and firms within their jurisdictions when correcting for resource misallocations. Thus, they simply follow the normative decision rules derived under the assumption of a single government. There is no need for the grant-in-aid, because no government need be directly concerned with any other jurisdictions. In our view, this is yet another reason for rejecting the traditional model of optimal federalism. It seems implausible that intergovernmental relations would be of no consequence in a federalist system of governments, even under first-best assumptions.

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00028-5

Our alternative model of federalism, presented in Chapter 27, defined the social welfare optimum as an equilibrium in which each government maximized its own dynastic social welfare function, with the restriction that the arguments of each government's social welfare function are the social welfare functions of those governments immediately below it in the fiscal hierarchy. Grants-in-aid are required in this model to resolve the distribution question, since all but the lowest level governments must tax and transfer resources lump sum among the governments immediately below them in the fiscal hierarchy. In the parlance of grants-in-aid, these lump-sum grants would be unconditional, nonmatching, and closed-ended: unconditional, because one government cannot dictate to any other government how to dispose of the funds, the "states' rights" criterion; nonmatching and closed-ended, because the interpersonal equity conditions require straight resource transfers of some finite amount. Notice, too, that the "grants" are negative for those governments that must surrender resources.

Our alternative model shares with the conventional model the attribute that grants-in-aid are not required for allocational purposes in a first-best policy environment with a perfect correspondence of local functions. Simultaneously with satisfying all possible interpersonal equity conditions, satisfying all necessary pareto-optimal conditions proceeds government-by-government in the usual manner. To develop a further role for grants-in-aid, then, requires introducing some second-best distortion into the policy environment.

Second-Best Policy Environment

Imperfect Correspondence

A second-best restriction commonly analyzed in the literature is a maintained imperfect correspondence for an externality-generating activity, which causes each local government to follow the wrong decision rule. Imagine the following situation.[1] Community A, consisting of H_A individuals, provides a Samuelsonian nonexclusive public good in amount $\overline{X}_G$, the services of which are consumed directly by its own citizens. In determining the amount $\overline{X}_G$, the government of A follows the standard first-best decision rule:

$$\sum_{h_A=1}^{H_A} \mathrm{MRS}^{h_A}_{X_G,X_{h_A 1}} = \mathrm{MRT}_{X_G,X_1} \qquad (28.1)$$

where X_1 is a private good.

Suppose that H_B citizens of contiguous community B benefit from the existence of X_G in community A even though they cannot directly consume the services of X_G. For example, X_G may be police protection that has the spillover effect of reducing criminal activity in community B. In effect, then, X_G in community A becomes an aggregate external economy for the citizens of community B, entering into each person's utility function. The aggregate gain to community B's citizens on the margin can be represented as

$$\sum_{h_B=1}^{H_B} \mathrm{MRS}^{h_B}_{X_G,X_{h_B 1}}$$

with each MRS^{h_B} measured positively. The true first-best pareto-optimal conditions are therefore

$$\sum_{h_A=1}^{H_A} \mathrm{MRS}^{h_A}_{X_G,X_{h_A 1}} + \sum_{h_B=1}^{H_B} \mathrm{MRS}^{h_B}_{X_G,X_{h_B 1}} = \mathrm{MRT}_{X_G,X_1} \qquad (28.2)$$

Without any intervention from a higher level government in the fiscal hierarchy, X_G will be misallocated (presumably undersupplied), because community A ignores the second set of terms on the left-hand side of Eqn (28.2). The situation exemplifies the notion of an imperfect correspondence, since the jurisdictional boundaries of community A, which makes the allocational decision on X_G, do not encompass all citizens affected by the production and consumption of X_G.

There is no need for a grant-in-aid in this case. The next highest government in the fiscal hierarchy, one that includes the citizens of both A and B, could provide X_G to the citizens of A in accordance with Eqn (28.2). It does have the option, however, of allowing community A to decide on the level of X_G as before and influencing its decision with an appropriate grant-in-aid. Hence, its choice is fully analogous to that in single-government models of aggregate externalities, in which the government can either dictate the consumption of the good or use a Pigovian subsidy and maintain decentralization. A society committed to federalism would presumably choose the grant-in-aid since it promotes decentralized local autonomy, much as a single government under capitalism would choose decentralized subsidies for aggregate externalities.

As discussed in Chapter 6, the appropriate subsidy is a per-unit subsidy, equal to the aggregate gain to the citizens of B on the margin, or

$$s = \sum_{h_B=1}^{H_B} \mathrm{MRS}^{h_B}_{X_G,X_{h_B 1}}$$

which, in this case, is a grant-in-aid from the higher level government to community A. The grant, depicted in Fig. 28.1, would be conditional, matching, and

1. A similar example appears in Oates (1972), pp. 95–104. Oates's Chapter 3 and appendices provide an excellent analysis of the uses of grants-in-aid within the conventional model of fiscal federalism.

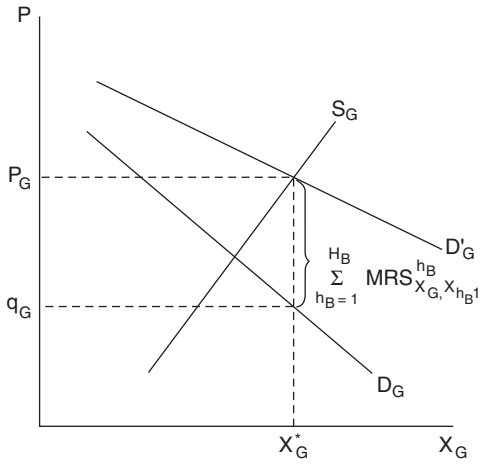

FIGURE 28.1

open-ended: conditional on expenditures for X_G with a matching rate equal to the ratio s/P_G at the optimum, where P_G is the producer price of X_G (see Fig. 28.1), and open-ended because it is not optimal to limit the size of the grant to any value other than $s \cdot X^*_G$, where X^*_G is determined by the receiving government.

These simple grant-in-aid examples can be quite misleading, however. Localities tend to provide the same kinds of public services, so that the actual pattern of externalities is likely to be far more complex than depicted in our simple story. If community *A*'s police expenditures generate external economies in community *B* (and, possibly, other neighboring communities), then community *B*'s police expenditures can be expected to generate external economies for all its neighbors, including *A*. But, if this is so, then the spillover component of the externality is likely to be individualized by community, in which case the required pattern of grants-in-aid becomes extremely complex.

To see the possibilities, define $(X_{G_A}, X_{G_B}, \ldots, X_{G_C})$ as the vector of the individual community outputs, *C* in number, with $X_G = X_{G_A} + X_{G_B} + \ldots + X_{G_C}$ the aggregate output of X_G across all communities. If the aggregate X_G enters each person's utility function, then a single matching grant is appropriate, with $s = \sum_{\text{all } h} \text{MRS}^h_{X_G, X_{h1}}$. Referring again to police expenditures, the assumption is that the spillover effects on criminal activity within a region depend upon aggregate police expenditures across all communities within the region.

Although expenditures on police may give rise to an aggregate externality, each community is more likely to receive the most benefit from police expenditures in its contiguous communities and increasingly less benefit from police expenditures in ever more distant towns. If so, then the spillover externality remains individualized and pareto optimality requires a complex set of subsidies, one for each town. Moreover, the subsidies are interdependent, with each matching rate dependent upon police expenditures in every community. Thus, the situation is exactly analogous to the case of individualized externalities arising from private sector activities.

We have seen that aggregate externalities admit to relatively simple solutions, whereas individualized externalities do not. The existence of federalism, with imperfect correspondences, adds nothing to the complexity of the problem. Even if the next highest government in the fiscal hierarchy chose to provide X_G, it would still follow the same decision rules, providing, of course, that the direct services of each individual X_{G_i} are consumed exclusively by members of the corresponding community, as posited in our example. If there is not even a perfect correspondence for the direct consumption of these services, then a set of grants-in-aid is unlikely to be appropriate. In this case, the next highest government should decide upon the level of the aggregate $\overline{X}_G$ and its individual subcomponents. For instance, police services may be exclusive by town because the laws of each town forbid police to cross-jurisdictions. But if there really is an imperfect correspondence here, then these exclusions are arbitrary and nonoptimal. Fewer, larger police departments with a regional orientation would be the optimal solution, but these would have to be provided by the next highest government in the fiscal hierarchy.

Note finally that the analysis carries through in both the conventional model of federalism in which only the national government has a social welfare function, or in our alternative model in which each government possesses a social welfare function. So long as income is optimally distributed according to the interpersonal equity conditions of each model, allocational issues dichotomize from distributional concerns just as in single-government models.

These same points apply to externalities generated by private sector activity. Unless the direct component of the activity can be localized within a single community (say, a production externality arising at a particular site), grants-in-aid are unlikely to be pareto optimal. And even if pareto optimality could be achieved by the grants-in-aid, it may not be the most direct fiscal tool. Not surprisingly, grants-in-aid are most appropriate for publicly provided services.

Consider the example of a production site located in community *A*. Suppose its external diseconomies affect both citizens in *A* and those in other neighboring towns. If town *A* taxes the producer, it will undoubtedly base the tax on the marginal damages only to its own citizens. The next highest government could design a negative conditional matching grant (i.e., a tax) levied on town *A* that would optimally adjust for the broadened scope of the external diseconomy, but an additional direct tax on the producer would seem less cumbersome. Other more complex situations, such as the individualized pollution example in

which production at multiple sites along a river generates external diseconomies for the other firms, can best be solved by producer taxes established by a higher level government and not by a set of grants-in-aid to a number of localities. There is no compelling reason to involve lower level governments as intermediaries in correcting for private sector externalities.

ALTERNATIVE DESIGN CRITERIA

That actual grants-in-aid bear little relationship to theoretical design criteria is hardly surprising, because the theory is so difficult to apply in this instance. In terms of our alternative model, distributional norms based on social welfare functions can never be more than suggestive to the policy maker. In terms of imperfect correspondences for externality-generating public services, varying matching formulas across "local" governments on the basis of marginal external benefit or harm may be unconstitutional. Faced with these realities, economists have resorted to developing practical design criteria that are at least roughly consistent with the underlying theory.

A surprising feature of the more practical literature is that it has tended to focus on distributional concerns, more in line with our alternative model of optimal distribution under federalism than with the mainstream position. A principal question is how to design grants to correct for perceived resource imbalances either across states (for federal grants) or across localities (for state grants). This focus makes sense at a practical level because many federal and state grants in the United States do attempt to direct aid disproportionately toward poorer states and localities. Examples are the federal grants to support states' public assistance payments under Temporary Assistance for Needy Families (TANF) and Medicaid and state grants to support local public school expenditures.

The LeGrand Guidelines

In the mid-1970s, Julian LeGrand suggested three sensible practical guidelines for grant-in-aid programs whose goals are redistributional (LeGrand, 1975). First, the grants must be a function of the real income or wealth of the receiving government, commonly referred to as its fiscal capacity. LeGrand argues that jurisdictions with fiscal capacities below some target level should receive aid and jurisdictions above the target should pay a tax (receive a negative grant). In contrast, existing grant-in-aid programs always give something to all governments. The political motivations behind giving something to everyone are clear, but such grants tend by their very nature to have limited redistributional power. Note, also, that fiscal capacity accounts for differences in prices across communities, the relative expenditures required to achieve comparable levels of public services.

LeGrand's second guideline is that the amount of aid received (tax paid) should be independent of any expenditure decisions made by the receiving government. This guideline honors two principles: Redistributional policy ought to properly be concerned with each government's overall initial level of resources, and, consistent with the federalist ideal, the grantor should not attempt to influence the specific spending decisions of lower level governments.

LeGrand's third guideline states that grants should vary directly with the receiving government's fiscal effort, the idea being that governments with less interest in providing public services should receive correspondingly less aid. This criterion is somewhat troublesome because it tends to contradict the second guideline. It implies that the grantor will try to influence the overall level of public services beyond the giving or taking of resources, although not the composition of these services. In any case, it is a commonly accepted principle. The US Congress has frequently incorporated effort parameters into aid formulas.[2]

LeGrand shows that basing grants-in-aid on differences in fiscal capacity automatically incorporates each community's fiscal effort. To see this, let

T_i = total taxes per capita collected by government i
P_i = a price index of public services provided by government i
E_i = the effective tax rate in government i, the effort parameter
Y_i = the per capita tax base in government i

The fiscal capacity of government i is Y_i/P_i. LeGrand defines a purchasing power effort (PPE) ratio as

$$\text{PPE}_i = \frac{T_i}{E_i P_i} \tag{28.3}$$

where purchasing power refers to the purchasing power of the taxes. But $T_i = E_i Y_i$. Therefore,

$$\text{PPE}_i = \frac{T_i}{E_i P_i} = \frac{E_i Y_i}{E_i P_i} = \frac{Y_i}{P_i} \tag{28.4}$$

LeGrand's PPE ratio is the same as fiscal capacity.

Under LeGrand's preferred grant-in-aid formula, the grantor picks a target PPE ratio or fiscal capacity, $\text{PPE}_\text{T} = Y_\text{T}/P_\text{T}$. The per capita grant, G_i, is then designed to put all jurisdictions at that target PPE_T. Thus, G_i is such that

$$\frac{T_i + G_i}{E_i P_i} = \frac{Y_i}{P_i} + \frac{G_i}{E_i P_i} = \frac{Y_\text{T}}{P_\text{T}} \tag{28.5}$$

2. When Congress replaced Aid to Families with Dependent Children (AFDC) with TANF, it stipulated that the states could not reduce the expenditures on public assistance that they had been making under AFDC.

or

$$G_i = E_i\left(\frac{P_i}{P_T}Y_T - Y_i\right) \tag{28.6}$$

The grant received (tax paid) depends upon a locality's fiscal effort as embodied in the tax rate, and its relative fiscal capacity, defined as the difference between its per capita tax base and the target per capita tax base adjusted by the differences in the prices of public services in the locality relative to the target community. Hence, all three of LeGrand's criteria are satisfied by this simple formula.

LeGrand's formula would lead to a substantial amount of redistribution, since richer than average towns would actually pay taxes. By including E_i, the formula also addresses a problem with federalism that many people find particularly inequitable, namely, wealthy communities can offer better public services than the poorest communities even though their tax rates are only a fraction of the tax rates in the poorest communities. LeGrand's formula doubly rewards the poor communities who have high tax rates. Finally, if one concedes that social welfare rankings may properly be functions of fiscal effort, among other things, this simple formula is reasonably consistent with the redistributional decision rules of our alternative model of fiscal federalism. It bears roughly the same relationship to these norms as the Haig–Simons ability-to-pay criterion does to the interpersonal equity conditions of single-government social welfare maximization. Both substitute income for utility, although the Haig–Simons criterion contains nothing comparable to the fiscal effort term.

Applying LeGrand's Principles: Bradbury et al.

LeGrand's grant formula is much too egalitarian to be politically acceptable. A more practical version of his proposal would be to close only a portion of the disparities in fiscal capacity:

$$G_i = kE_i\left(\frac{P_i}{P_T}Y_T - Y_i\right) \quad k < 1 \tag{28.7}$$

subject to the constraints

$$G_i \geq 0 \quad \text{all } i \tag{28.8}$$

and

$$\sum_i G_i N_i = D \tag{28.9}$$

where D is the budget given to the distributional granting authority for the grants to reduce fiscal disparities (N_i is the population of locality i). Equation (28.8) ensures that no communities with fiscal capacities greater than Y_T/P_T would be taxed under the formula. The granting authority would maintain the budget constraint by varying k and the reference community Y_T. A high Y_T combined with a low k gives smaller amounts of aid to more communities, and vice versa. Taxes to support the grants would come from general tax revenues, not from levies on the high-fiscal-capacity communities.

In the early 1980s, Katherine Bradbury et al. were commissioned by the Massachusetts state government to design an equalizing grant program for distributing 5% of the state's grant budget, approximately $110 million, to the cities and towns with low fiscal capacities (Bradbury et al., 1984). They approached the problem in the spirit of LeGrand, but they used a different measure of fiscal disparity in the aid formula. They based their formula on what they termed a community's fiscal gap, equal to

$$\text{Gap}_i = \overline{E}C_i - \overline{t}B_i \tag{28.10}$$

where

$\overline{E}$ = the average per capita expenditures across all communities
C_i = the cost of providing the average expenditures in community i
$\overline{t}$ = the average tax rate across all communities
B_i = the per capita tax base in community i.

In other words, a community's fiscal gap is the difference between what it would have to spend to provide the average local public service bundle and the tax revenues it would raise if it applied the average tax rate across all communities to its tax base.

A reference, or target, fiscal gap is defined in the same way:

$$\text{Gap}* = \overline{E}C_T - \overline{t}B_T \tag{28.11}$$

The grant formula closes a portion of the difference between a community's fiscal gap and the reference fiscal gap,

$$\begin{aligned} A_i &= k\left(\text{Gap}_i - \text{Gap}_T\right) \\ &= k\left[\overline{E}\left(C_i - C_T\right) - \overline{t}\left(B_i - B_T\right)\right] \quad A_i \geq 0 \end{aligned} \tag{28.12}$$

where A_i is the per capita grant. The first term on the right-hand side (RHS) is the cost disadvantage suffered by community i relative to the reference community, and the second term is community i's tax-base disadvantage. The main deviation from LeGrand's principles is that the Bradbury et al. formula does not include an effort term.

Bradbury et al. argue that the average expenditure level $\overline{E}$ and the cost of providing the services C_i should be based on regression analysis. They also believe that the relative cost advantages or disadvantages should reflect only environmental factors that are beyond the immediate control of the communities, such as population density, the condition

of the housing stock, and the crime rate. They posit a supply of expenditures function:

$$E_i = E_i\left(\overrightarrow{\mathbf{S}}_i, \overrightarrow{\mathbf{P}}_i, \overrightarrow{\mathbf{C}}_i\right) \tag{28.13}$$

where

$\overrightarrow{\mathbf{S}}_i$ = the vector of public services offered in community i
$\overrightarrow{\mathbf{P}}_i$ = the vector of input prices for the factors used to produce the public service vector in community i
$\overrightarrow{\mathbf{C}}_i$ = the vector of environmental factors that influence the cost of providing the public service in community i.

The demand side of the model is a standard median voter model (described later) in which the median household solves an as-if maximization problem in terms of a numeraire private composite commodity and the vector of public services, subject to its individual budget constraints and the overall community budget constraint. The supply relationship, Eqn (28.13), enters as the expenditures in the overall community budget constraint. The analysis leads to a reduced-form equation for overall public expenditures (individual public service outputs are not measurable):

$$E_i = f\left(\overrightarrow{\mathbf{V}}_i, \overrightarrow{\mathbf{A}}_i, \overrightarrow{\mathbf{P}}_i, \overrightarrow{\mathbf{D}}_i, \overrightarrow{\mathbf{C}}_i\right) \tag{28.14}$$

where

$\overrightarrow{\mathbf{V}}_i$ = the average (mean) property value in community i
$\overrightarrow{\mathbf{A}}_i$ = a vector of other resources available to community i, such as other grants-in-aid
$\overrightarrow{\mathbf{D}}_i$ = a vector of taste parameters, "demand" factors.

The demand factors Bradbury et al. chose were per capita income and the percentage of the population ≥65 years. The five environmental cost factors were population density, the condition of the housing stock, the ratio of children in the public schools to the entire population, the crime rate, and the poverty rate. They had no data on variation of input prices, $\overrightarrow{\mathbf{P}}$, across the cities and towns. Equation (28.14) was estimated on a sample of 300 Massachusetts towns.

To estimate $\overline{E}$ in the grant formula, Eqn (28.12), they set the values of all the explanatory variables in Eqn (28.14) equal to their average values across all 336 cities and towns. To compute the relative cost term C_i in their grant formula, they estimated $\hat{E}_i$ by setting the values of all the explanatory variables except $\overrightarrow{\mathbf{C}}$ at their average values, and the values of the variables $\overrightarrow{\mathbf{C}}$ at their actual values in community i. Then $C_i = \hat{E}_i/\overline{E}$ or $C_i\overline{E} = \hat{E}_i$ in the grant formula.

In applying the Bradbury et al. formula, the state

1. Set the reference $\text{Gap}_T = 0$, to maximize the number of communities receiving aid.
2. Set an additional condition that every community receives a grant of at least $5 per capita from the budget set aside for these grants.
3. Defined the fiscal gaps to include existing state aid, $\overline{A}$:

$$\text{Gap}_i = \overline{E}C_i - \overline{t}B_i - \overline{A}_i \tag{28.15}$$

Finally, since all grant, expenditure, and tax-base variables are in per capita terms, the cities and towns received a proportion of the entire distribution budget equal to the product of their fiscal gaps and population divided by the sum of the fiscal gaps times populations of all the aided localities.

Bradbury et al. proposed that the aid be adjusted each year using the same estimating equation for E_i and just adjusting the values of the explanatory variables.

The EU Cohesion Grants

A major grant program very much in the spirit of LeGrand's principle of fiscal equalization is the European cohesion grants, which constitute about 35% of the European Union's budget. A primary motivating factor in the formation of the European Union was to reduce fiscal disparities throughout the member nations, a goal the European Union refers to as convergence. The cohesion grants are the principal means to this end. They are targeted to regions within countries whose per capita gross national income is substantially below the overall EU average (less than 90% of the average under one of the grant programs, and less than 75% of the average under two other grant programs that distribute the majority of the funds). The grants have a number of goals, but over 60% of the total grant funds are specifically to support convergence. The main difference from the LeGrand prescription is that the cohesion grants are project grants, not unconditional grants. They are targeted to projects such as infrastructure investment, business investments, and job training in the poorer regions. Grant recipients propose specific projects and are expected to pay part of the costs.[3]

Redistributing through Matching Grants

As our final example of practical grant design criteria, we will consider Martin Feldstein's proposal for remedying unequal local public educational expenditures (Feldstein, 1975). In the early and mid-1970s, a number of state supreme courts ruled that financing public educational expenditures entirely from local property taxes was inherently discriminatory, since wealthier communities could

3. Information about the cohesion grants is available in the annual EU financial reports. See, for example, EU Budget 2012.

provide better education with less fiscal effort, that is, lower tax rates.[4] The states were required to design a more equitable statewide financial arrangement that would somehow provide transfers from the wealthier to the poorer communities. Feldstein reasoned that the courts' decisions imply a fiscal solution that sets the elasticity of educational output with respect to wealth equal to zero ($E_{\text{Ed,w}} = 0$). He suggested using a matching grant for this purpose, in which the matching rate applied to any one community is inversely proportional to its wealth. To achieve this goal, one needs reliable econometric estimates of the price and income (wealth) elasticities of educational expenditures independent of a new grant program. These estimates can then be used to design the required matching rates.

To see how this would work, suppose it is possible to estimate a constant elasticity demand-for-education equation across communities of the form

$$\text{Ed} = CP^{\alpha}W^{\beta} \tag{28.16}$$

where

Ed = a measure of educational output per capita
P = the price of a unit of educational output
W = a measure of per capita community wealth
α, β = the price and wealth elasticities
C = a constant term embodying all other factors influencing the demand for education.

Rewriting Eqn (28.16) in log form:

$$\log \text{Ed} = C' + \alpha \log P + \beta \log W \tag{28.17}$$

Next, define a matching aid formula that makes the net-of-aid price a function of wealth according to the constant elasticity form

$$P = W^{k} \tag{28.18}$$

or

$$\log P = k \log W \tag{28.19}$$

where

k = the elasticity of the net price with respect to wealth.

Substituting Eqn (28.19) into Eqn (28.17) yields

$$\begin{aligned} \log \text{Ed} &= C' + \alpha(k \log W) + \beta \log W \\ &= C' + (\alpha k + \beta)\log W \end{aligned} \tag{28.20}$$

With this matching program,

$$\frac{\partial \log \text{Ed}}{\partial \log W} = E_{\text{Ed,W}} = \alpha k + \beta \tag{28.21}$$

Setting $E_{\text{Ed,w}} = 0$ implies

$$k = -\beta/\alpha \tag{28.22}$$

Thus, the required matching rate elasticity just equals the ratio of the wealth and price elasticities of education within the state, at least for a log-linear demand for education function. Feldstein estimated an education equation for a cross-section of Massachusetts communities to demonstrate his technique. The required matching rate elasticity for Massachusetts turned out to be between 0.33 and 0.37 (Feldstein, 1975, p. 85).

It is worth repeating that matching grants for which the matching rate varies with respect to income or wealth have no role in the first-best theory of federalism and are at best only suggested by second-best considerations. Nonetheless, if the law requires neutralizing the effect of wealth on educational opportunity within states, then Feldstein's grant-in-aid formula provides a direct way of achieving this goal.

ESTIMATING THE DEMAND FOR STATE AND LOCAL PUBLIC SERVICES

The final issue in our analysis of grants-in-aid is an empirical one—the response of receiving governments to grants-in-aid. We have to begin, however, with a detour on the modeling of state and local governments' demands for the services that they offer. The reason is simply that any empirical analysis of how state and local governments respond to grants-in-aid is naturally embedded in a model of their demand for public services.

A number of different demand models exist in the literature, but it is fair to say that the median voter model has emerged over the past 25 years as the favored empirical model for estimating the determinants of state and local spending decisions. Its only serious competitor is the qualitative response model based on surveys that ask people questions about their desired increases or decreases in public spending. Therefore, we begin with a discussion of the median voter and survey models before turning to the response to grants-in-aid. Grants-in-aid will then be analyzed in the contest of the median voter model because it is by far the more widely used model.

The Median Voter Model

That the median voter model became so popular is testimony to the difficulties that economists face in trying to model state and local governments. In truth, the median voter model is highly problematic. It rests on extremely strong political and economic assumptions that are unlikely to hold for almost any state or locality. It also encounters some econometric problems that were not recognized in the earlier literature. These weaknesses notwithstanding, the

4. *Serrano v. Priest* in California was the landmark decision. Refer to Serrano v. Priest.

median voter model is the predominant model for estimating the spending decisions of state and local governments.

The Political Assumptions

The model's name derives from its political assumptions. It is motivated by the direct democracy, one-person–one-vote, small town meeting, in which the citizens congregate periodically to discuss and vote on government spending and tax issues. A simple majority determines the outcome of the vote: A proposition wins if it gains 50% of the votes plus one. Under these rules, the preferences of the median voter are decisive so long as the preferences of the citizens are monotonic over the issue being voted on.

Panel (a) of Fig. 28.2 illustrates the case of a local public good G. The line G indicates the most preferred level of G for each citizen on the horizontal axis, ordered by their preferences for G. Assume an odd number of citizens so that the median voter is identified. G_{median} is the amount of G preferred by the median voter. Consider any $G < G_{median}$ such as G_1. In a vote for G_1 against a small increase in G, $G_1 + \Delta G$, the majority prefer $G_1 + \Delta G$. Similarly, for any $G > G_{median}$ such as G_2, the majority prefer a small decrease in G, $G_2 - \Delta G$. G_{median} is the only amount of G that can command the required 50% plus one-vote majority against the next larger or smaller amount of G. The preferences of the median voter are decisive.

The idea that the median voter's preferences will be decisive in actual elections is difficult to accept for a number of reasons. First, the result relies crucially on preferences for public goods and services being monotonic. If, instead, the preferences for G are as pictured in panel (b) of Fig. 28.2, then G_1 wins a simple-majority election. The median voter prefers G_{max}, which cannot win a simple majority. More generally, democratic decision making runs up against Arrow's impossibility theorem. We showed in Chapter 4 that social preferences over a distinct set of choices can cycle and fail to establish a clear winner if individual preferences over the choices are not single peaked.

Cycling is almost certain to occur if people are asked to vote for choices that contain a bundle of two or more public goods, as is often the case. Town meeting members typically vote on entire budgets that include a variety of services: education, transportation, recreation, public safety, and so forth. Panel (c) of Fig. 28.2 gives one example in which people are asked to vote for a combination of two public goods, G_1 and G_2. The points 1, 2, and 3 indicate the most preferred bundles for persons 1, 2, and 3, respectively.

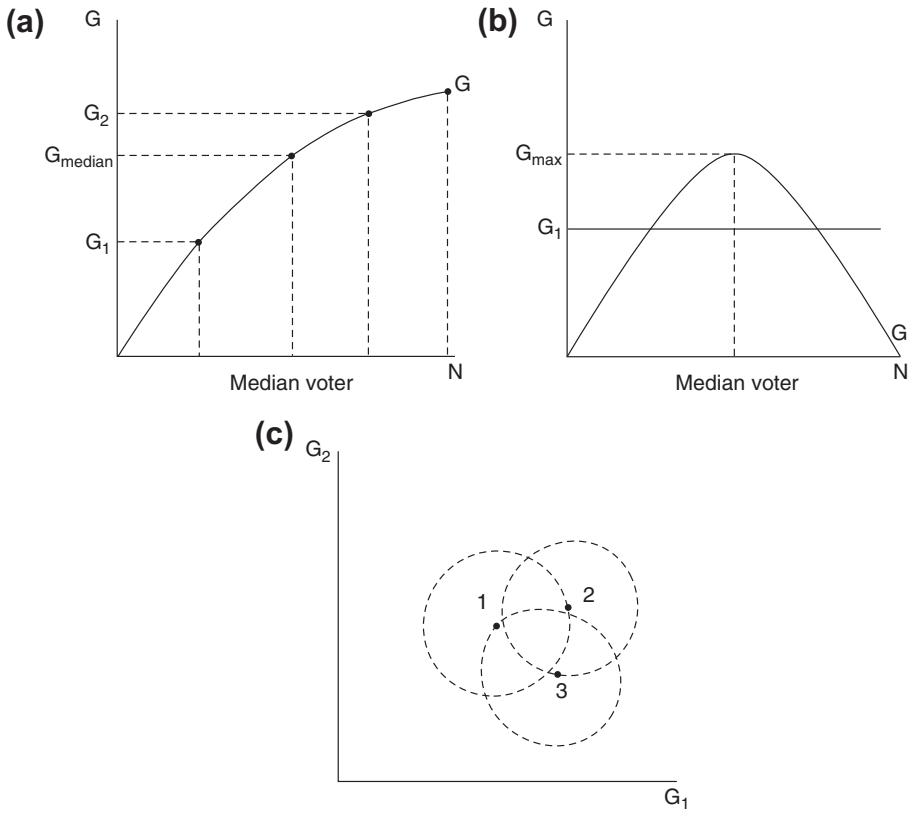

FIGURE 28.2

Utility declines with the distance from the most preferred choice in every direction, so that the indifference curves radiate from the most preferred choice in a circle-like pattern. In the example, the individual preferences for the three choices are as follows:

Person 1: 1 P 2 P 3
Person 2: 2 P 3 P 1
Person 3: 3 P 1 P 2

These are the same preferences as in Chapter 4, with the same intransitive results: Two of the three people prefer 1 to 2, 2 to 3, and 3 to 1. No clear winner emerges.

Add the political realities of representative democracy, political parties vying for votes, special interest groups lobbying legislators and members of the executive branch, and self-interested bureaucrats, and the idea that the median voter's preferences are likely to be decisive on any issue is problematic in the extreme. The median voter model can hardly be descriptive of any government except the very small towns that might still use the democratic town meeting to decide public issues. Nonetheless, researchers have used the median voter model as the basis for estimating even states' decisions.

The Economic Assumptions

The fundamental economic assumption of the model is the standard one in the federalism literature, that each household solves an as-if maximization problem to determine its most preferred level of public services.[5] The as-if maximization leads to an estimating equation that is then expanded to test a number of propositions about state and local public services of interest to economists. The model also requires a number of additional assumptions so that it can be estimated using readily available data. One of these assumptions ties the model to the political assumption that the median voter is decisive.[6]

As-If Maximization

Begin in the usual manner. Each household i has a fixed endowment of income, Y_i, that is spent on two goods, X_i and G_i. X_i is a private composite commodity and serves as the numeraire. G_i is the locally provided public good, which is purchased at price p_i. The household's as-if maximization problem is

$$\max_{(X_i,\, G_i)} U^i(X_i, G_i)$$
$$\text{s.t. } Y_i = X_i + p_i G_i$$

The resulting demand curve for the household's preferred G_i leads to the basic estimating equation (in log-linear form)

$$\ln G_i^* = a + b \ln Y_i + c \ln p_i + e_i \tag{28.23}$$

where e_i is the error term.

The price p_i requires some explanation because households pay taxes and let the government buy the good on their behalf. They do not buy the public good directly. Consequently, the effective price depends on how the government collects taxes to pay for G. Assume that the model applies to localities that use a property tax. Then p_i turns out to be the product of two terms: the ratio of the value of household i's property, V_i, to the total property value in the locality, V, (V_i/V), and the supply price of G, equal to q. This follows because household i pays a property tax equal to tV_i, where t is the property tax rate. Multiplying and dividing by V, the tax is

$$tV_i = tV(V_i/V) \tag{28.24}$$

But the local government budget constraint requires that total taxes equal total expenditures, or

$$tV = qG \tag{28.25}$$

Substituting Eqn (28.25) into Eqn (28.24) yields

$$tV_i = (V_i/V)qG \tag{28.26}$$

so that $p_i = (V_i/V)q$. Substituting for p_i in Eqn (28.23) yields the basic estimating equation:

$$\ln G_i^* = a + b \ln Y_i + c \ln\left(V_i/V\right) + c \ln q + e_i \tag{28.27}$$

Equation (28.27) is typically expanded and adjusted in the following ways to produce the final estimating equation.

Allowing for Congestion

The first adjustment is to allow for the possibility of congestion in the public good. Write:

$$G_i^* = G_i/N^\alpha \quad \alpha = (0, 1) \tag{28.28}$$

where G_i is the actual G to be provided by the government. $\alpha = 0$ is the nonexclusive good; each household receives the full services of G. $\alpha = 1$ is the purely private good; it takes N units of G to provide one unit of G to household i. Expressing congestion in this way implies that the effective price of G to household i is

$$p_i^* = p_i N^\alpha \tag{28.29}$$

For example, household i has to pay for N units of G to get one effective unit of G if G is purely private; therefore, the effective price is N times the price as given by

5. The household is the appropriate economic decision-making unit in the median voter model.
6. The discussion in this section follows that of Atkinson and Stiglitz in Atkinson and Stiglitz (1980).

Eqn (28.23). Expressing Eqns (28.28) and (28.29) in natural logs,

$$\ln G_i^* = \ln G_i - \alpha \ln N \tag{28.30}$$

$$\ln p_i^* = \ln p_i + \alpha \ln N \tag{28.31}$$

Substituting Eqns (28.31) and (28.30) in Eqn (28.27) and rearranging terms yield the congestion-adjusted basic estimating equation:

$$\begin{aligned}\ln G_i &= a + b \ln Y_i \\ &= c \ln(V_i/V) + c \ln q + \alpha(1+c)\ln N + e_i\end{aligned} \tag{28.32}$$

Expenditures per Person

Governments routinely publish data on expenditures rather than separate series on prices and outputs of their services. Indeed, defining the output of a school system or the police force is somewhat ambiguous and open to interpretation. Outputs are ambiguous, so too are prices. Therefore, the dependent variable in the estimating equation is typically expenditures per capita for each category of public services. Write:

$$E_i/N = (qG_i)/N \tag{28.33}$$

or

$$\ln(E_i/N) = \ln q + \ln G_i - \ln N \tag{28.34}$$

Substituting for ln G in Eqn (28.32) and rearranging terms yields

$$\begin{aligned}\ln(E_i/N) &= a + b \ln Y_i + c \ln(V_i/V) \\ &\quad + (1+c)\ln q + [\alpha(1+c) - 1]\ln N + e_i\end{aligned} \tag{28.35}$$

Other Determinants of G

The individual utility functions differ by a set of taste parameters, $\vec{\mathbf{Z}}$, that reflect such things as differences in household composition and size, the households' inherent interest in supporting public education or public safety, and neighborhood characteristics such as population density and proximity to a major city. These parameters are simply added to the estimating equation (expressed here in log form):

$$\begin{aligned}\ln(E_i/N) &= a + b \ln Y_i + c \ln(V_i/V) + (1+c)\ln q \\ &\quad + [\alpha(1+c) - 1]\ln N + \vec{\mathbf{d}} \ln \vec{\mathbf{Z}} + e_i\end{aligned} \tag{28.36}$$

The Supply Price q

The supply price q presents a problem in estimating Eqn (28.35) because it is generally not observable. State and local governments do publish wage and salary data for the employees in each service category, however. Therefore, researchers typically substitute wages and salaries for the supply price by assuming that production is constant returns to scale and least cost efficient and that the market for capital is national in scope.

To see the implication of these assumptions, assume that G is produced with capital (K) and labor (L) according to the CRS production function:

$$G = f(K, L) \tag{28.37}$$

If production is least cost efficient, then

$$f_K/f_L = r/w \tag{28.38}$$

where r is the cost of capital and w is the wage. Furthermore, the expansion path of capital and labor is linear given CRS. Therefore, if (K^*, L^*) is the optimal combination of capital and labor to produce G^*, then (λK^*, λL^*) is also an optimal combination of capital and labor and produces $\lambda G^*[\lambda G^* = f(\lambda K^*, \lambda L^*)]$. But the total cost of production is

$$\text{TC} = rK + wL \tag{28.39}$$

Therefore,

$$\lambda\text{TC} = r(\lambda K) + w(\lambda L) \tag{28.40}$$

Since scaling capital and labor by λ scales both G and TC by λ for any r and w, the total cost function for G has the form

$$\text{TC} = h(r, w)G \tag{28.41}$$

Hence, marginal cost is

$$\text{MC} = h(r, w) \tag{28.42}$$

Finally, assume that r is set in the national market and that the supply price, q, is equal to (or at least proportional to) marginal cost. Then q is proportional to the w, which varies across localities. This is the justification for substituting wages for q in the estimating equation, Eqn (28.36), yielding the final estimating equation:

$$\begin{aligned}\ln(E_i/N) &= a + b \ln Y_i + c \ln(V_i/V) + (1+c)\ln w \\ &\quad + [\alpha(1+c) - 1]\ln N + \vec{\mathbf{d}} \ln \vec{\mathbf{Z}} + e_i\end{aligned} \tag{28.43}$$

Whose Equation?

An equation such as Eqn (28.43) applies to each household as the solution to its as-if maximization problem. Each locality, however, supplies only one level of G (E/N) for each public service. The question, then, is whose G (E/N) obtains in the locality, and the answer is that of the median voter. But then what income and property value does the

median voter have? The answer again is that the median voter resides in the household with the median income and the median property value. That is, the median voter has the median value of all the economic variables that vary by households. This heroic assumption is necessary because only median values of these variables are routinely available in the local (and state) Census of Governments. It permits estimation of the model with a cross-section of data on localities (states).

An implication of this assumption is that preferences for public services are monotonic in incomes and property values. If not, then the median voter in terms of preferences for G will not necessarily have the median income or property value. One common test of this assumption is whether the full elasticity of G with respect to Y is positive (assuming public goods are normal goods). Let $(V_i/V) = t_i$ and write:

$$dG/dY = \partial G/\partial Y + (\partial G/\partial t_i)(\partial t_i/\partial Y) \qquad (28.44)$$

The first term is the direct effect of Y on G, and the second term is the indirect effect that works through the effect of Y on the value of the property (house) that people buy. To convert to elasticities, multiply all terms in Eqn (28.44) by Y/G and multiply the second term on the RHS by (t_i/t_i), yielding

$$E_{G,Y}(\text{full}) = E_{G,Y}(\text{direct}) + E_{G,t_i} \cdot E_{t_i,Y} \qquad (28.45)$$

or

$$E_{G,Y}(\text{full}) = b + cE_{t_i,Y} \qquad (28.46)$$

from Eqn (28.43). $E_{t_i,Y}$ is the elasticity of the value of property (housing) with respect to income.

Further Conceptual Difficulties

The estimating model faces a number of additional difficulties even given all its heroic assumptions. A brief list would include the following.

Renters

The model essentially ignores renters by focusing on housing values. The median voter might not have anywhere close to the median house value when a locality contains a substantial percentage of renters, especially if renters have different demands for public services than homeowners.

Nonvoters

The model assumes that everyone votes. This is a terrible assumption for the United States, especially at the state and local level. The propensity to vote is known to be directly related to income and education. Consequently, the median voter is likely to have income and property value well above the median.

Commercial and Industrial Property Taxes

The model should probably be adjusted for localities that levy taxes on commercial and industrial property as well as residential property, but it is not clear how. Expanding the tax base has the direct effect of lowering t_i. At the same time, however, commercial and industrial establishments increase the demand for certain kinds of public services. Therefore, the effective t_i of the median voter could rise or fall. What matters in any event is the median voter's perception of how commercial and industrial property taxation affects his or her t_i, whatever the truth might be.

Multiple-Service Budgets

The fact that votes are often taken on entire budgets strains the belief that voters can attach effective prices to each of the individual services. Do they really see the supply prices of the various services? If not, do they assume they are simply proportional to wages, and do they know the wages of the teachers, police, firefighters, sanitation workers, and so forth? More generally, do they know their share of the total property value in the community?

The Results

All these difficulties notwithstanding, the models have generated two results with a fair degree of consistency. One is that income elasticities exceed price elasticities (in absolute value), and often by a considerable margin. One survey of the median voter literature concluded that the estimated income elasticity appears to be on the order of 2/3 and the estimated price elasticity somewhere between $-1/4$ and $-1/2$.[7] These estimates are roughly consistent with elasticities estimated using models other than the median voter model. The finding of very low price elasticities for local (and state) public services is the rule in the empirical federalism literature. The income and price elasticities are such that the full elasticity of public services with respect to income, Eqn (28.46), is positive, as expected.

The second consistent finding is that α is close to one for many public services; they appear to be much more like private goods than nonexclusive goods. Wallace Oates thinks this result may be an illusion, however. He believes that as populations increase in communities, many public services expand in discrete steps and become more complex, essentially different kinds of services, as illustrated in Fig. 28.3. If Oates is correct, then the public service could be nonexclusive within each step, as pictured. But the estimation will fit a line through the steps with a considerable slope (the dotted line), falsely suggesting a high degree of congestion (a high α) (Oates, 1988).

7. Bergstrom et al. (1982), p. 1199 and Table IV, p. 1200.

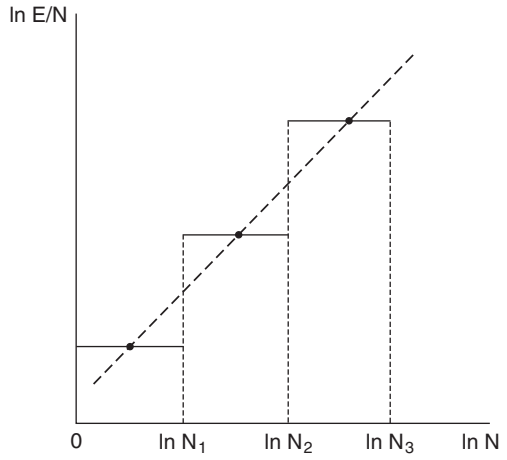

FIGURE 28.3

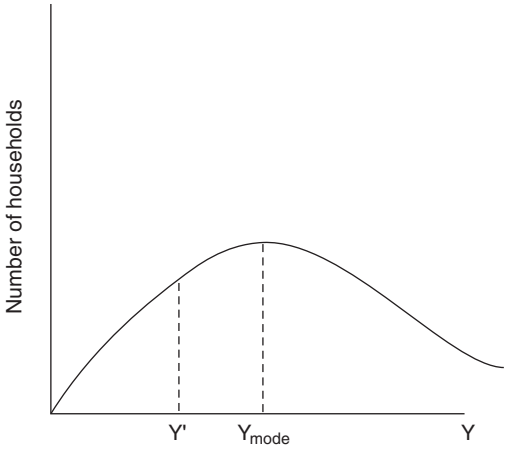

FIGURE 28.4

Econometric Problems: Tiebout Bias

Estimating an equation such as Eqn (28.43) with data on median incomes and property values across communities is almost certain to lead to biased estimates of these coefficients. The bias occurs because households select communities on the basis of the public services in each community. For this reason, the bias is referred to as Tiebout bias. Tiebout bias is a specific instance of the estimation bias that results whenever economic agents select options on the basis of the dependent variable. Most of the existing empirical literature based on the median voter model makes no attempt to correct for Tiebout bias, certainly none before the mid-1980s. We will illustrate the bias with respect to income.[8]

Suppose that households solve their as-if maximization problems and there are enough towns so that all the households are able to find their most preferred G. The matches are perfect. In this case, the estimation of G would proceed as it does for private goods and services. Collect data on a random sample of individuals throughout the geographic region and estimate the equation (expressed here in level form):

$$G_i = a + bY_i + e_i \tag{28.47}$$

Assume the error term e_i is normally distributed with mean zero and is uncorrelated with the independent variables. The estimate of b would be an unbiased estimate of the true b.

The matches are far from perfect, however, and this leads to the bias, as follows. Suppose the distribution of income throughout the entire region is unimodal as pictured in Fig. 28.4, and select an income Y' less than Y_{mode}. Households with income Y' have a distribution of tastes for G given by the error term in Eqn (28.47). Refer to Fig. 28.5. When $e = 0$, the number of households who want G' is f_0. G' is the level of G that corresponds to Y' according to the true relationship between G and Y.

Now consider the distribution of tastes for G by households with incomes equidistant from Y', $Y' - \delta$ and $Y' + \delta$. Refer to Fig. 28.6. G' is the preferred level of G for f_1 households with incomes $Y' - \delta$, and G' is the preferred level of G for f_2 households with incomes $Y' + \delta$. Given the unimodal distribution of Y throughout the region, f_2 is larger than f_1, as pictured. Since households are selecting communities on the basis of G, the median income of the community that offers G' is greater than Y'.

Similarly, select an income Y'' greater than the mode and consider the distribution of preferences for G″, the G associated with Y'' according to the true relationship between G and Y. By the same argument as above, the median income of the community that offers G'' is less than Y''.

The implication is that the estimated relationship between G and Y based on Y_{median} across localities is biased

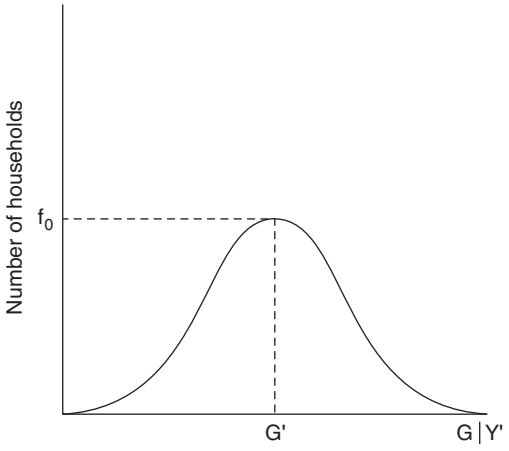

FIGURE 28.5

8. The seminal article on Tiebout bias is Goldstein and Pauly (1981), pp. 131–144.

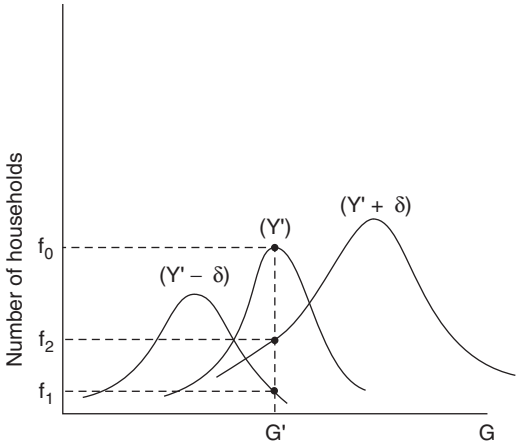

FIGURE 28.6

upward, as pictured in Fig. 28.7. The true relationship between G and Y is the solid line; the estimated relationship based on the median voter model is the dotted line.

Given the positive bias, the question arises whether the true income elasticities for public services really are much greater than the price elasticities, as is so commonly reported in the literature. Also, the receipt of exogenous grants-in-aid would be treated much like the receipt of income by the median voter. Therefore, estimates of the response to grants-in-aid may be biased upward as well.

Surveys

Economists also use surveys of individuals to estimate demand functions for state and local expenditures. The use of surveys was made possible by the development of the econometric theory for estimating qualitative response models.

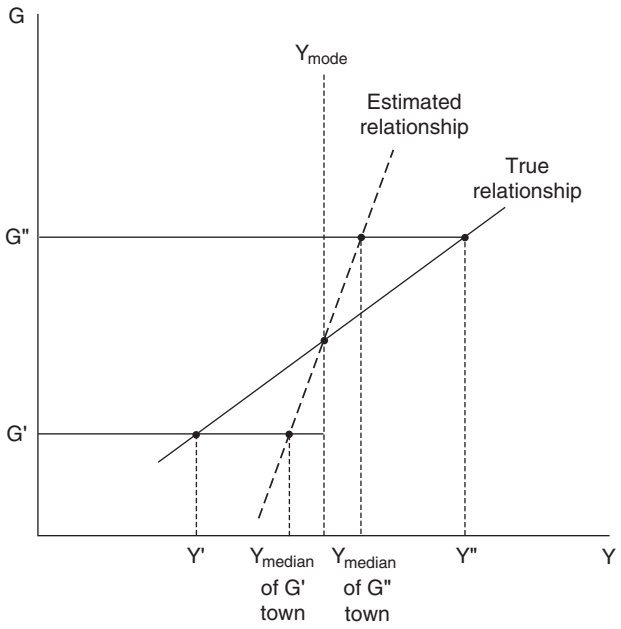

FIGURE 28.7

The survey approach has two distinct advantages over the median voter model. One is that it avoids many of the unrealistic political and economic assumptions required to estimate the median voter model. Another is that it allows the researcher to collect detailed information on individual characteristics that are likely to affect households' demands for particular public services, such as the education and employment status of the head of household, the age composition of the household, race, ethnicity, the number of children in the public schools, and so forth. The principal disadvantage of the survey approach is concern about the reliability of the survey responses. Economists are inherently skeptical about surveys, having been trained to observe what people do rather than what they say. Surveys of large numbers of people are also quite time consuming and costly.

One of the earliest, and still best known, of the survey studies was by Theodore Bergstrom, Daniel Rubinfeld, and Perry Shapiro in the early 1980s (Bergstrom et al., 1982). We will use their study as an example of the survey approach.

Bergstrom, Rubinfeld, and Shapiro conducted a survey of 2001 people in Michigan to try to determine their demand for local school expenditures. They simply asked if people desired "more," "the same," or "less" spending. If respondents said "more," the surveyor noted that this would require higher taxes and asked if they still preferred more spending. If the answer was no, their responses were considered to be "the same." Twenty-five percent of the respondents said they would prefer "more" spending; 58%, "the same," and 17%, "less."

To develop an estimating model, Bergstrom, Rubinfeld, and Shapiro adopted the standard assumption that people solve an as-if maximization problem to determine their desired spending levels for public services, which leads to a demand equation of the general form

$$\ln G_i^* = a + b \ln Y_i + c \ln p_i + \vec{\mathbf{d}} \ln \vec{\mathbf{Z}}_i + \ln e_i \quad (28.48)$$

where G_i^* refers to the desired level of expenditures per pupil on public education rather than the output. The estimating equation mimics the estimating equation of the median voter model, other than containing a much richer vector of personal characteristics, $\vec{\mathbf{Z}}$. For example, the relevant price p_i is the product of the respondent's tax share of an increased dollar of expenditure on education and the supply price of education. The supply price is proxied by the ratio of average teacher salaries to average salaries of all workers in the county in which the respondent resides. The distribution of $\ln e_i$ is assumed to be logistic because the estimating

strategy is based on the probabilities of the responses, as explained below.

To see how the estimating model is developed, think of the demand function, Eqn (28.48), as the sum of its deterministic portion, labeled ln $D(X_i)$, and the error term ln e_i. Also, assume for the moment that there are only two categories: "more" spending and "less" spending. Finally, let G^A stand for the actual level of spending per pupil in a respondent's community.

The respondents will presumably say "more" if $G_i^* > G^A$ and "less" if $G_i^* < G^A$. The dependent variable G_i^* is unobserved, and the observed responses can only take on the two values "more" and "less," which can be represented as 1 ("more") and 0 ("less"). The limitation of the observed responses to 0 and 1 suggests a probabilistic interpretation of the model. The estimation framework should have the property that the probability of a "more" response increases the larger the [ln $D(X_i)$ − ln G^A], and approaches 1 as [ln $D(X_i)$ − ln G^A] → ∞. Conversely, the probability of a "less" response increases the larger the [ln G^A − ln $D(X_i)$] and approaches 0 as [ln G^A − ln $D(X_i)$] → ∞

Consider the "more" response. $G_i^* > G^A$ implies that

$$\ln D(X_i) + \ln e_i > \ln G^A \tag{28.49}$$

or

$$\ln e_i > \ln G^A - \ln D(X_i) \tag{28.50}$$

Giving Eqn (28.50) a probabilistic interpretation, the probability that the respondent will say more is the Pr(ln e_i) > ln G^A − ln $D(X_i)$. But, ln e_i has a logistic distribution, which is symmetric. Therefore, an equivalent statement is: The probability that the respondent will say "more" is the Pr(ln e_i) < ln $D(X_i)$ − ln G^A. But, this is just the value of the logistic cumulative density function evaluated at [ln $D(X_i)$ − ln G^A], F(ln $D(X_i)$ − ln G^A), as pictured in Fig. 28.8. Similarly, the probability that the respondent will say "less" is 1 − F(ln $D(X_i)$ − ln G^A).

The coefficients of Eqn (28.48) can be estimated by maximizing the likelihood function of the "more" and "less" responses expressed in terms of the binomial distribution (each response is considered to be one draw from the distribution):

$$\max L = \prod_{i=\text{more}} F\left(\ln D(X_i) - \ln G^A\right) \prod_{j=\text{less}} \left(1 - F\left(\ln D(X_j) - \ln G^A\right)\right) \tag{28.51}$$

The Threshold Effect

The response "the same" is accounted for by adding a threshold parameter, $\delta > 1$, such that the respondent replies

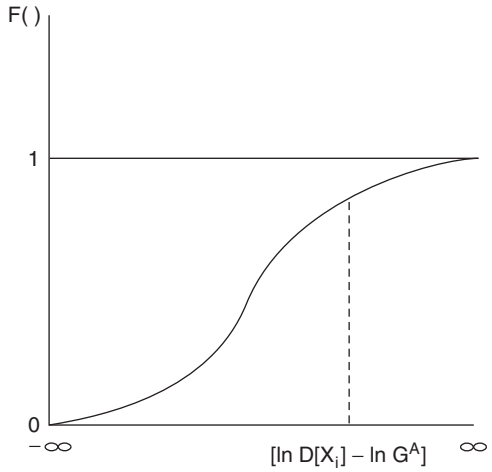

FIGURE 28.8

"more" if $G_i^* > \delta G^A$ and "less" if $G_i^* < G^A/\delta$. This leads to three regions under the logistic cumulative density function:

$$\begin{aligned}
&\Pr(\text{"more"}): \Pr(\ln e_i) < \ln D(X_i) - \ln G^A - \ln \delta \\
&= F(\ln D(X_i) - \ln G^A - \ln \delta) \\
&\Pr(\text{"less"}): \Pr(\ln e_i) > \ln D(X_i) - \ln G^A + \ln \delta \\
&= 1 - F(\ln D(X_i) - \ln G^A + \ln \delta) \\
&\Pr(\text{"the same"}): 1 - \Pr(\text{"more"}) - \Pr(\text{"less"}) \; F(\ln D(X_i) \\
&-(\ln G^A + \ln \delta) - F\left(\ln D(X_i) - \ln G^A - \ln \delta\right)
\end{aligned}$$

The estimation maximizes a multinomial likelihood function with the three sets of product terms defined over the three responses.

The Results

Bergstrom, Rubinfeld, and Shapiro found a number of interesting results. One was that the estimated income and price elasticities, $E_y = 0.64$ and $E_p = -0.39$, were in the range of the estimates from the studies using the more aggregated median voter model. Another was that the ability of the survey approach to capture a broad set of individual characteristics was an important advantage. They found that respondents were more likely to want more spending on education if they were black, Jewish, renters, elderly, a school employee, or had children in the public schools. Conversely, respondents were more likely to want less spending if they were unemployed, retired, disabled, or sent their children to private schools. Finally, the estimated threshold effect was large, 1.5, suggesting perhaps that Tiebout searching leads to reasonably good matches of people's desired spending on public education.

Tiebout Bias

The survey approach is subject to Tiebout bias because the respondents selected their communities partly on the basis of the dependent variable. Bergstrom, Rubinfeld, and Shapiro did not account for Tiebout bias in their original study, but Rubinfeld, Shapiro, and Judith Roberts published a follow-up study 5 years later on the same survey data that did try to account for the Tiebout bias (Rubinfeld et al., 1987). Their results indicate that correcting the estimation for the bias is important. In particular, the new estimates produced very low income and price elasticities, both on the order of 0.1 (in absolute value). The very low income elasticity is really quite unusual in the local public sector empirical literature. The threshold effect also increased, to 1.65. In other words, spending on education would have to be 65% above or below its existing level before people would want less or more spending. They appear to be quite satisfied with the status quo, suggesting even more successful Tiebout matching of preferences than the earlier study.

THE RESPONSE TO GRANTS-IN-AID

The literature on the response to grants-in-aid has been motivated by two factors: (1) grants-in-aid have long been very important to state and local governments, and (2) governments receive many different kinds of grants-in-aid. Recall that grant formulas vary across three dimensions: conditional—unconditional, matching—nonmatching, and closed-ended—open-ended. Public sector economists have had a natural theoretical interest in the expected responses to the various possible combinations of formula parameters. Should it matter, for example, whether governments receive conditional or unconditional grants, matching or nonmatching grants, and so forth? On an empirical level, econometric analysis has tried to pinpoint the actual response to existing grants, both for its own sake and as a test of the theoretical analysis. Taken together, this body of literature is as extensive as any in public sector analysis, yet both the theoretical and empirical analyses of grant response have been far from conclusive.

The Flypaper Effect

The most consistent result in the empirical literature is that governments' responses to exogenous grants-in-aid far exceed their responses to exogenous increases in other resources, most particularly increases in the total wealth or income within the state or locality. Fernando Aragon reports that empirical studies in the United States estimate that, on average, the marginal propensity to spend out of increases in grants is 0.64, whereas the marginal propensity to spend out of equal increases in income is on the order of 0.05—0.10.[9] This discrepancy has been termed the flypaper effect, because grant funds appear to "stick where they hit."

The empirical analysis leading to the finding of a flypaper effect rests on two fundamental principles. The first, noted earlier, is that a model of how governments respond to grants-in-aid should be part of the same model used to determine their demands for the various public services. Since the median voter model is the favored model for estimating the demand for state and local services, it is also the model used to estimate the response to grants-in-aid. This implies that the response of the median voter to the government's receipt of a grant determines how the government itself will respond to the grant, since the median voter is decisive. Consequently, the standard approach simply adds the grant parameters to the usual estimating equation of the median voter model. We will consider a truncated version of the full model that highlights the price and income terms, and write the basic estimating equation for spending category i to be adjusted for grants-in-aid as

$$\ln(E/N)_i = a + b \ln Y_{\text{med}} + c \ln t_{\text{med}} + (1+c)\ln q_i + \ldots + e_i \tag{28.52}$$

where

$(E/N)_i$ = expenditures per capita on category i
$t_{\text{med}} = (V_{\text{med}}/V)$ = the tax share of the median household, the ratio of its property value to the total property value in the community
q_i = the supply price of spending category i, proxied in the estimation by the wages and salaries of the employees in category i.

The second principle is that the median voter should respond to a grant-in-aid exactly as a consumer would respond to an individual transfer payment. The operative question is how the median household perceives that the receipt of a grant-in-aid affects its own budget constraint. Assuming that the median household spends an endowment of income, Y_{med}, on a private numeraire good X and public good G_i, the median voter's budget constraint is

$$X + p_{\text{med}} G_i = Y_{\text{med}} \tag{28.53}$$

or

$$X + t_{\text{med}} q_i G_i = Y_{\text{med}} \tag{28.54}$$

The various transfer (grant) possibilities were discussed in Chapter 10 in the context of pareto-optimal redistributions. To review: If the median household's community receives an open-ended matching grant for G_i at a matching rate of m, the relevant net price of G_i to the

9. Aragon (2008). For a summary of the empirical literature see Hines and Thaler (1995), pp. 217—226.

median household becomes $P_{med}(1 - m)$. If the grant is closed-ended with a grant limit of A, the following possibilities arise:

1. Unconditional, closed-ended grant that can be spent on any good—Grants of this form are equivalent to increases in income to the recipient.
2. Conditional, closed-ended grant targeted to good i—As long as the receiving government spends more on good i than the amount of the grant, the grant is equivalent to an unconditional grant, that is, to an increase in income. The recipient can undo the conditions of the grant by adjusting its expenditures from its own resources on the aided and unaided items. The condition that the funds be spent on good i matters only if the recipient spends none of its own resources on good i, in which case, it is forced to a corner solution on its budget constraint. This condition is virtually never satisfied for grants-in-aid and certainly not for any of the major grants. The receiving government always spends more than the maximum amount of the conditional grant under any of the major grants, such as the federal highway grants, the public assistance TANF grants, and the state education grants.
3. Conditional, closed-ended matching grant—This grant is also exogenous and equivalent to an unconditional grant so long as the receiving government reaches the limit of the aid. The matching rate is irrelevant beyond the limit; every additional dollar requires a dollar of funds from the locality's own resources. The grant acts as a price-reducing matching grant if the recipient remains within the matching region, which is unlikely. The possibilities suggest that all (important) grants other than open-ended matching grants should be equivalent to unconditional exogenous grants. Conditioning or targeting the grant to a specific type of expenditure should not matter. Therefore, an exogenous grant of A represents an increase in the resources of the median voter equal to $t_{med}A$, the household's share of the grant funds. The median household's budget constraint under an exogenous grant is

$$X + p_{med}G_i = Y_{med} + t_{med}A \tag{28.55}$$

Almost all federal grants are closed-ended conditional grants, matching or nonmatching, that are equivalent to unconditional grants. The only exception is Medicaid, which is an open-ended matching grant to the states that reimburses them from 50% to 83% for whatever expenditures they incur under Medicaid, with the matching rates inversely related to state income. Most state grants to localities are also closed-ended conditional grants that are equivalent to unconditional grants. Therefore, researchers typically aggregate all exogenous grant funds and add them to the estimating equation on a per capita basis. The grant-adjusted estimating equation incorporating open-ended matching and exogenous grants is

$$\ln(E/N)_i = a + b\ln Y_{med} + c\ln t_{med}(1 - m) + (1 + c)\ln q_i + \ldots + f\ln(A/N) + e_i \tag{28.56}$$

where (A/N) equals total exogenous grants per capita, not just those targeted to good i.

The flypaper effect associated with the exogenous grants involves a comparison of the coefficient estimates of b and f. The exact comparison depends on how the median voter views (A/N).

Since the median voter views the grant A as equivalent to an increase in its resources of $t_{med}A$, the expectation is that[10]

$$\partial G_i/\partial(t_{med}A) = \partial G_i/\partial Y_{med} \tag{28.57}$$

To represent the grants on a per capita basis, define $t^*_{med} = Nt_{med}$. Then

$$\partial G_i/\partial(t^*_{med}A/N) = \partial G_i/\partial Y_{med} \tag{28.58}$$

If $t^*_{med} = 1$, then the median household treats the receipt of an exogenous per capita grant as equivalent to an increase in income. But t^*_{med} is likely to be much less than 1 (equivalently, t_{med} is likely to be much less than $1/N$), for two reasons. First, the distributions of income and most forms of wealth such as property values are almost always highly skewed toward the high end, as indicated in Fig. 28.9, such that the mean income or wealth is well above the median income or wealth. Hence, $t_{med} = (V_{med}/V)$ is likely to be less than $1/N$ in all states and localities; thus, $t^*_{med} < 1$. In addition, localities are able to export some of their property tax burden to citizens outside the locality,

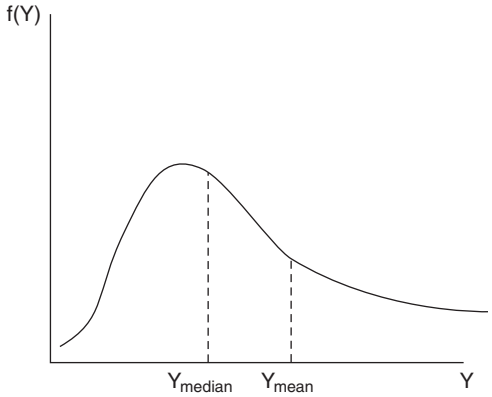

FIGURE 28.9

10. Expressing the flypaper effect in terms of output instead of per capita expenditure is less cumbersome. Only the output component of E/N changes as income or grants change.

which lowers the median voter's tax share even more. Estimates of the proportion of exported local tax revenues in the United States range from 0.65 to 0.85. To test for a flypaper effect, therefore, rewrite Eqn (28.58) as

$$\partial G_i/\partial(A/N) = t^*_{\text{med}}\partial G_i/\partial Y_{\text{med}} \tag{28.59}$$

Convert Eqn (28.59) into elasticities consistent with the log-linear form of the estimating equation, Eqn (28.56), by multiplying both sides by $[Y_{\text{med}}(A/N)/G]$:

$$Y_{\text{med}}E_{G,(A/N)} = t^*_{\text{med}}(A/N)E_{G,Y_{\text{med}}} \tag{28.60}$$

$$f = [t^*_{\text{med}}(A/N)/Y_{\text{med}}]b \tag{28.61}$$

Equation (28.61) indicates the expected equivalence in the median household's response to the community's receipt of a per capita grant or to an increase in its income. A flypaper effect exists if $f > [t^*_{\text{med}}(A/N)/Y_{\text{med}}]b$. Since both t^*_{med} and $(A/N)/Y_{\text{med}}$ are less than one, f has to be greater than only a fraction of b to generate a flypaper effect. In fact, most estimates of equations such as Eqn (28.56) find that $f > b$; the flypaper effect appears to be very large. Governments respond much more to grants-in-aid than to an equivalent increase in income, that is, much more than the theory would predict. A strong flypaper effect is also typically found in estimating models other than the median voter model.

Possible Explanations of the Flypaper Effect

Economists have offered a number of possible explanations for the strong flypaper effect, all of them based on the idea that the grant-adjusted median voter model is somehow misspecified or incomplete. We will briefly consider two possibilities: fiscal illusion and the partial equilibrium nature of the model.

Fiscal Illusion

Wallace Oates believes that voters may suffer from a form of fiscal illusion. They may view exogenous grants as reducing the price of public services as well as increasing community resources because they confuse average and marginal prices. The average price of public services is the ratio of total tax collections to total expenditures (expressed as a vector):

$$P_{\text{avg}} = \left(\frac{\sum_{i=1}^{N} T_i}{q\cdot G}\right) = \left(\frac{q\cdot G - A}{q\cdot G}\right) \tag{28.62}$$

The perceived price effect introduces a bogus substitution effect that could help to explain the extra kick to public spending that grants appear to have. This explanation suffers from the normal skepticism of any theory based on people falling victim to illusions. In addition, the estimated price elasticities of local public goods are quite small, perhaps too small to explain the rather large flypaper effect, even if voters do suffer from average price illusion. At best, average price illusion appears to be only a partial explanation.

Combining Grant and Tax Effects

Ronald Fisher (Fisher, 1982) has introduced a more promising explanation in our view, one that has not been adequately tested to date. Fisher argues that the response to grants in the median voter model as described above is incomplete, in effect a partial equilibrium rather than a general equilibrium analysis. It views voters as savvy enough to see how the receipt of a grant by their community affects their own budget constraints in their as-if maximization problems. But if they are savvy enough to see this, then they are also savvy enough to realize that the grants have to be paid for by taxes collected from the higher level granting government. They would realize, in other words, that the net increase in resources to the community and to themselves is the grant less the taxes paid to finance the grant. Given the progressivity of the federal personal income tax, a grant to a high-income community from the federal government could easily represent a net decrease in resources to the community and the median voter.

In any event, to determine the effect of the grant on its budget constraint, the median household should properly compare its share of the local property tax, t_{med}, with its share of federal taxes paid within the community to finance the grant nationwide, $t^{\text{fed}}_{\text{med}}$. The total response to the grant is

$$dG = (\partial G/\partial t_{\text{med}}A)d(t_{\text{med}}A) - (\partial G/\partial Y_{\text{med}})d(t^{\text{fed}}_{\text{med}}T^{\text{fed}}) \tag{28.63}$$

where T^{fed} are the total personal income taxes collected from the community to finance the grant program. Equation (28.63) indicates that a resource-neutral grant program from the community's point of view, $A = T^{\text{fed}}$, will not be resource neutral from the median household's point of view unless $t_{\text{med}} = t^{\text{fed}}_{\text{med}}$.

Fisher's general point is that the researcher has to get the model right before attempting to describe and estimate a flypaper effect.[11]

11. Holsley (1993), was one of the first economists to take a general equilibrium approach and consider the financing of the grants in analyzing the response to grants-in-aid. She found that grants to local education generated both price and income illusion among voters, with the income illusion resulting from misunderstanding that taxes have to be paid to the donor government to support the grant program.

The Deadweight Loss of Local Taxes

Jonathan Hamilton pointed out that the flypaper effect does not have to be due to fiscal illusions or other behavioral anomalies. It could simply be the result of the deadweight efficiency costs of raising local taxes that rational consumers take into consideration when responding to increases in incomes or grants.

To illustrate, Hamilton develops a simple model in which identical consumers within a locality have a source of lump-sum income I and receive utility from a private good X and another good Y provided by the local government. The goods are defined such that their prices are both equal to one. The local government collects taxes T_L and receives grants from the federal or state government equal to G, with both taxes and grants measured per person. The cost of the taxes to the consumer is given by the function $g(T_L)$, specified in terms of units of X, with $g' > 1$ and $g'' > 0$. $g' > 1$ implies that the local tax is a distorting tax: a dollar of tax revenue has more than a dollar cost to the consumer because of the deadweight loss of the tax, and the consumer understands this. $g'' > 0$ implies that the deadweight loss increases at an increasing rate. In contrast, the grants G are assumed not to involve any efficiency cost. Hamilton thinks it is appropriate to ignore Fisher's point that rational consumers should consider the federal taxes raised to finance the grants, and any inefficiencies associated with them, since a consumer can safely assume that the amount of grants he receives is independent of the federal taxes he pays. In any event, he sees this as the spirit in which the flypaper effect is tested.

The consumer's utility function is $U = U(X,Y)$ and the budget constraint is $I = X + g(T_L)$. The government's budget constraint is $Y = T_L + G$. Using the consumer's budget constraint to substitute for X in the utility function and the government's budget constraint to substitute for Y, the government's problem is to

$$\text{Max } U = U(I - g(T_L),\ (T_L + G)$$
$$\text{w.r.t. } T_L$$

The FOC is

$$-U_1 g' + U_2 = 0 \tag{28.64}$$

The flypaper effect compares the responses $\frac{dY}{dI}$ and $\frac{dY}{dG}$. Notice, though, that $\frac{dY}{dI} = \frac{dT_L}{dI}$ and $\frac{dY}{dG} = 1 + \frac{dT_L}{dG}$ from the government's budget constraint. Therefore, totally differentiate the FOC to compute $\frac{dT_L}{dI}$ and $\frac{dT_L}{dG}$.

$$\left(U_{11}(g')^2 - 2U_{12}g' + U_{22} - U_1 g''\right)dT_L = \left(U_{11}g' - U_{12}\right)dI \tag{28.65}$$

or

$$\frac{dY}{dI} = \frac{dT_L}{dI} = \frac{U_{11}g' - U_{12}}{U_{11}(g')^2 - 2U_{12}g' + U_{22} - U_1 g''} = \frac{U_{11}g' - U_{12}}{D} \tag{28.66}$$

D is assumed to be negative since it is the second derivative of U with respect to the government's decision variable T_L. Therefore, the numerator of Eqn (28.66) is also assumed to be negative since the publicly provided good Y is assumed to be a normal good $\left(\frac{dY}{dI} > 0\right)$.

Similarly,

$$\frac{dT_L}{dG} = \frac{U_{12}g' - U_{22}}{D} \tag{28.67}$$

Adding $\frac{D}{D}$ to both sides of Eqn (28.67) yields

$$\frac{dY}{dG} = 1 + \frac{dT_L}{dG} = \frac{U_{11}(g')^2 - U_{12}g' - U_1 g''}{D} \tag{28.68}$$

The flypaper effect is the difference

$$\frac{dY}{dI} - \frac{dY}{dG} = \frac{(U_{11}g' - U_{12})(1 - g') + U_1 g''}{D} \tag{28.69}$$

The first two terms in the numerator are negative by the assumptions that Y is a normal good and that raising taxes generates a deadweight loss ($g' > 1$). The last term is positive under the assumption that deadweight loss is increasing in tax revenue. Since D is negative, $\frac{dY}{dI} - \frac{dY}{dG} < 0$, which is the flypaper effect. Public good provision responds more to grants than to equal increases in income (Hamilton, 1986).

More recently, Fernando Aragon developed a variation of Hamilton's model in which he assumed that the consumer's utility function was quasi-linear ($U = X + H(Y)$) and that the costs of raising taxes took the form of administrative and compliance costs rather than deadweight losses. The quasi linearity of U allowed him to compute the size of the flypaper effect in terms of the difference between the local tax rate and the administrative costs of raising taxes. Since administrative costs are very low, the size of the flypaper effect is essentially equal to the level of the local tax rates in his model. The actual average local (and state) tax rates in the United States were high enough to account for almost all of the average size of the estimated flypaper effect.[12]

Project Grants and Bureaucrats

We conclude our discussion with the possibility of endogenous grant parameters. Most research on the

12. F. Aragon, *op. cit.*

response to grants-in-aid assumes that the grant parameters such as matching rates and grant limits are exogenous, equal to whatever the particular grant program describes them to be. This may not be true, however, if the bureaucrats administering the grants are pursuing their own agendas, in line with the public choice view of government officials. Howard Chernick pointed out long ago that the parameters of federal project grants to states and localities for such things as municipal waste treatment plants and community development initiatives were often subject to negotiation between the federal grant bureau and the potential recipients (Chernick, 1981). These grants are almost always matching grants with a spending limit, and potential recipients apply for aid on specific projects. Chernick noted that federal administrators are often willing to make a portion of the grant fungible so that it could be spent on anything, in return for the recipients accepting lower matching rates. The administrators do this because they want to maximize the number of projects funded by their limited grant budgets, and they accept the results of the empirical grants literature that income elasticities were much higher than price elasticities for state and local expenditures. Exploiting the higher income elasticities with partially fungible grants is a way to stretch their budgets.

Renegotiations of this kind present a problem for researchers in trying to estimate the responses to grants because the parameters of the grant are endogenous and other than what they appear to be. Figure 28.10 illustrates this. It analyzes a matching grant to finance expenditures on some educational project. The community purchases education (Ed) and all other public goods and services (O), and the supply prices of Ed and O are set equal to 1 for convenience along the community's budget constraint DG. The community moves to point C as a result of the grant, receiving a grant of FC and paying EF from its own budget. The announced matching rate is FC/EC, so that the assumed effective net of grant price is EF/EC = (DE/EC).

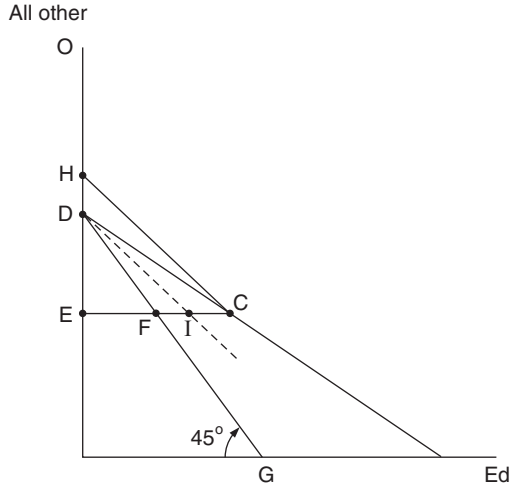

FIGURE 28.10

Suppose, in fact, the grant bureau and the recipient negotiated a deal in which the recipient received a fungible grant in amount HD in return for accepting a lower matching rate, a combination that also placed the recipient at C after the grant. The true effective net-of-grant price as a result of the negotiation is HE/EC. The total grant, G, is still FC, equal to the fungible portion IC (=HD) plus the matching portion FI.

Martin McGuire developed a procedure for estimating the responses to grants of this type (McGuire, 1975, 1979). It requires the assumption that the exogenous fungible portion, G_Y, equals a portion θ of the total grant:

$$G_Y = \theta G \tag{28.70}$$

This implies that the effective net-of-grant price, P = HE/EC, equals

$$P = \frac{\text{DE} + \theta G}{\text{EF} + \text{FC}} = \frac{E_{\text{own}} + \theta G}{E_{\text{own}} + G} \tag{28.71}$$

where E_{own} are expenditures by the community from its own resources. Adding and subtracting G in the numerator of Eqn (28.71), rearranging terms, and simplifying yields

$$P = 1 + \frac{(\theta - 1)G}{E_{\text{own}} + G} = 1 + (\theta - 1)m \tag{28.72}$$

where $m = G/(E_{\text{own}} + G)$ is the assumed, observed matching rate.

Therefore, an estimating equation on education that includes the independent variables

$$E = a(\text{exogenous grants}) + bP + \ldots \tag{28.73}$$

becomes

$$E = a\theta G + b[1 + (\theta - 1)m] + \ldots \tag{28.74}$$

where G and m are the observable total grant and matching rate. McGuire's estimating procedure allows for estimates of θ and the coefficients a and b. He finds that the price elasticities for these grants are negligible, thereby providing some justification for the federal administrators' willingness to negotiate.

The question remains why administrators in the recipient government have an incentive to negotiate in this manner. One possibility is that the receiving government happens to prefer having a portion of the grant with no strings attached, in line with the estimated income elasticities from the demand studies. Another possibility is more diabolical, in line with the public choice perspective that bureaucrats are aggressively self-serving. Suppose the local bureaucrats have private information about the negotiation and the project itself that they can hide from

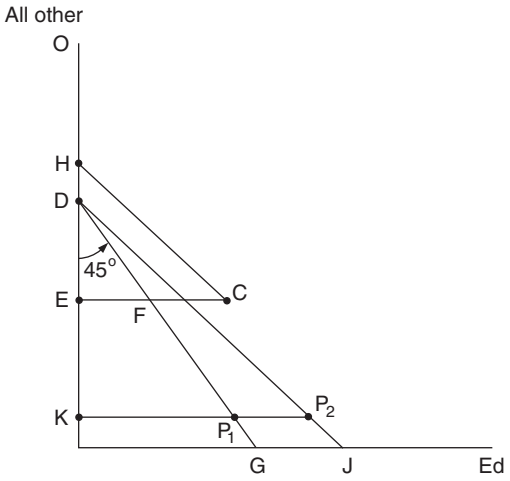

FIGURE 28.11

the legislature. They could then use their private information to their personal advantage as follows. Refer to Fig. 28.11.

Suppose, to begin with, that the local bureaucrats can hide the fungible portion of the grant from the legislature but not the entire grant. They tell the legislature that they were forced to accept the lower matching rate, but they do not mention the fungible portion. Therefore, they are pretending that the with-grant budget line is DJ, beginning from D but with the same slope (matching rate) as the true budget line beginning from H. A grant of FC on the true with-grant budget line is equal to a grant of P_1P_2 on the pretend with-grant budget line DJ. The bureaucrats tell the legislature that they are at point P_2 and require an amount KP_1 from the legislature to fund the recipient government's portion of the project. In fact, they only require EF from the legislature because they are really operating at C. Thus, they are able to pocket both the fungible portion of the grant plus the excess funds received from the government in the amount (KP_1 − EF). Whether the bureaucrat's information concerning public projects is really so private that they could get away with such schemes is doubtful, of course. But the story is indicative of the general point that private information is likely to provide local bureaucrats with incentives to negotiate with federal administrators on grant parameters.

REFERENCES

Aragon, F., November 2008. The Flypaper Effect Revisited. Discussion Paper. STICERD. London School of Economics.

Atkinson, A., Stiglitz, J., 1980. Lectures on Public Economics. McGraw-Hill, New York, pp. 322–326.

Bergstrom, T., Rubinfeld, D., Shapiro, P., September 1982. Micro-based estimates of demand functions for local school expenditures. Econometrica 50 (5), 1183–1205.

Bradbury, K., Ladd, H., Perrault, M., Reschovsky, A., Yinger, J., June 1984. State aid to offset fiscal disparities across communities. National Tax Journal 37 (2), 151–170.

Chernick, H., October 1981. Price discrimination and federal project grants. Public Finance Quarterly 9 (4), 371–394.

EU Budget 2012, Financial Report, Economic Commission, Publications Office of the European Union, Economic Cohesion for Growth and Employment, pp. 54–61.

Feldstein, M., March 1975. Wealth neutrality and local choice in public education. American Economic Review 65 (1), 75–89.

Fisher, R., November 1982. Income and grant effects on local expenditures: the flypaper effect and other difficulties. Journal of Urban Economics 12 (3), 324–345.

Goldstein, G., Pauly, M., October 1981. Tiebout bias on the demand for local public goods. Journal of Public Economics 16 (2), 131–144.

Hamilton, J., March 1986. The flypaper effect and the deadweight loss from taxation. Journal of Urban Economics 19 (2), 148–155.

Hines, J., Thaler, R., Fall 1995. Anomalies: the flypaper effect. Journal of Economic Perspectives 9 (4), 217–226.

Holsley, C., January 1993. Price and income distortions under separate spending and taxing decisions. Journal of Public Economics 50 (1), 93–114.

LeGrand, J., September 1975. Fiscal equity and central government grants to local authorities. Economic Journal 85 (339), 531–547.

McGuire, M., 1975. An econometric model of federal grants and local fiscal response. In: Inman, R., et al. (Eds.), Financing the New Fiscal Federalism. Johns Hopkins University Press for Resources for the Future, Inc., Baltimore, MD (Chapter 5).

McGuire, M., 1979. The analysis of federal grants into price and income components. In: Mieszkowski, P., Oakland, W. (Eds.), Fiscal Federalism and Grants in Aid. Urban Institute, Washington, D.C (Chapter 4).

Oates, W., 1972. Fiscal Federalism. Harcourt Brace Jovanovich, New York.

Oates, W., July 1988. On the measurement of congestion in the provision of local public goods. Journal of Urban Economics 24 (1), 85–94.

Rubinfeld, D., Shapiro, P., Roberts, J., August 1987. Tiebout bias and the demand for local public schooling. Review of Economics and Statistics 69 (3), 426–437.

Serrano v. Priest, L.A. 29820, Superior Court No. 938254.

Chapter 29

International Public Finance

Chapter Outline

Analysis of the international implications of public expenditure and tax policies began in earnest in the mid-1980s and has since become a major line of research in public sector economics. This is hardly surprising given the increasing globalization of the world's economies over the past 30 years. With the rapid growth of international trade in goods and services, the ever-increasing mobility of capital worldwide, and the rise of giant multinational enterprises (MNEs), the public sectors of all the industrialized market economies are ever more interdependent. The fiction of a closed economy, used throughout this textbook, is a useful device for developing the fundamental normative issues of public sector economics, but it is less useful as a framework for certain kinds of practical policy analysis. To give but one example, suppose the government in any one of the major industrialized nations raises the tax rates on its corporation income tax. An analysis of the effects of the tax increase on tax revenues and investment has to take into account the corporate income tax rates levied in the other major industrialized nations.

The literature on international tax and expenditures issues is now so large and so varied that a single chapter cannot begin to do it justice. All one can do is offer a highly selective introduction to some of the major issues, and highlight the more popular methods used to explore these issues in an international context. We focus on two of the main lines of research in the literature, both concerning taxation: (1) the taxation of mobile capital and (2) the ways in which MNEs can exploit differences in nations' tax policies to reduce their tax liabilities.

The taxation of mobile capital was the seminal issue that spawned the international public sector literature, dating from two articles published in 1986, one by George Zodrow and Peter Mieszkowski and the other by John Wilson.[1] These papers developed what quickly became the standard model for analyzing the taxation of internationally mobile capital, a model now commonly referred to as the ZMW model. The taxation of capital has remained as a focal point of theoretical research and public policy in international taxation. This is hardly surprising given its relevance. George Zodrow surveyed the associated empirical literature on the international mobility and taxation of capital, also a huge literature. An issue of central importance in empirical research has been determining the factors that influence foreign direct investment (FDI) by firms. Zodrow notes that FDI is important—capital is highly mobile across borders. In addition, researchers have found that FDI is very responsive to differences in corporate tax rates, with estimated elasticities as high as four, and the responsiveness appears to be increasing over time. The latest studies generally report the highest elasticities.

The analysis of MNEs is a more recent undertaking, one that requires a different modeling approach from the ZMW framework. A question of particular interest within this

1. Zodrow and Mieszkowski (1986), Wilson (1986). There are a number of excellent surveys of the international public finance literature that readers should consult who are seeking a broader introduction to the literature than this chapter can provide: Wilson (1999), Gordon and Hines (2002), Wilson and Wildasin (2004), and Keen and Konrad (2012) (the most analytical of the surveys in the style of a textbook chapter, but twice as long as our chapter and covering many more topics).

Public Finance. http://dx.doi.org/10.1016/B978-0-12-415834-4.00029-7

literature is why some countries try to attract MNEs by becoming tax havens with very low, even zero, tax rates on corporate income.

Here again Zodrow's survey of the empirical literature underscores the relevance of this line of research. He notes that competition for capital among countries has recently been shifting away from competition over tax rates toward the adoption of policies that make it easier for the MNEs to avoid paying taxes. And the evidence is overwhelming that MNEs have been able to avoid a considerable amount of tax liability by exploiting these policies, particularly those of the tax havens (Zodrow, 2010).

We begin with the taxation of mobile capital.

THE TAXATION OF MOBILE CAPITAL

Preliminaries: Normative Foundations

Concluding the textbook with a chapter on international tax issues is somewhat out of character because the thrust of this literature is overwhelmingly positive and practical in nature, not normative. This is certainly true of the analysis of tax shifting by MNEs and of the analysis of capital taxation as well. The corporation income taxes of all industrialized market economies are primarily source based, that is, they tax corporate profits earned in their own countries—at its source—regardless of whether the capital is owned by foreign or domestic firms. But there are residency-based features as well, because the income earned by the foreign-owned subsidiaries of MNEs is also taxed under the home country's corporation income tax, usually when the foreign profits are repatriated to the home country. This has the potential of producing a double taxation of corporate profits, a problem that countries avoid through bilateral tax treaties. The usual practice is for the home country to give its corporations a tax credit for any corporation income taxes paid abroad in the host countries.[2]

This is not the way to design an optimal tax system, either from a national or an international perspective. In the first place, a corporation income tax goes against the prescription of the Diamond–Mirrlees theorem in Chapter 24, that optimal commodity taxation implies production efficiency.[3] Since a separate tax levied on income from capital in the corporate sector but not in the unincorporated sector leads to an inefficient allocation of capital across the sectors, it cannot be part of an optimal tax system. Nor is it generally consistent with production efficiency to levy a source-based tax on income earned by foreign producers in the host country if they are subject to taxation again in their home country.

A better choice from a normative perspective would be to tax income from capital on a residency basis under a personal income tax, in which the residents of each country pay a tax on all income from capital they receive regardless of where the income was earned and whether it is earned in the corporate or unincorporated sectors. But tracking income earned abroad is often difficult and subject to the willingness of the tax collectors in the host countries to provide information on income earned to the home countries. Firms operating in foreign countries have an obvious incentive to hide income from the home tax authorities if they know that the host tax authorities are not providing the information. This is why countries choose to tax income from capital on a source basis.

The use of credits for foreign tax payments raises a separate issue concerning the appropriate point of view from a normative perspective. The usual assumption in normative analysis is that a country maximizes a social welfare function whose arguments are the utility functions of its citizens. The use of tax credits is inconsistent with this point of view.

The following simple model can be used to illustrate the point-of-view issue. Assume that producers use capital (K) and labor (L) to produce output according to the production function $f(K, L)$, and that the price of output is one. The income from capital subject to tax is $f(K, L) - wL$, where w is the wage, assumed to be the same worldwide. The overall supply of capital and labor is fixed within a country, referred to as the home country. Labor and capital are completely (costlessly) mobile across countries. Suppose firms in the home country have some incentive, which we need not identify, to send K^F of capital and L^F of labor abroad to produce in a host country and leave K^D of capital and L^D of labor at home to produce at home. The capital income earned abroad is $f(K^F, L^F) - wL^F$ and the capital income earned at home is $f(K^D, L^D) - wL^D$.

Each country levies an ad valorem source-based tax on income from capital, the host country at rate t^F and the home country at rate t^D.[4] The net of tax income from capital earned by the capital abroad after paying the foreign tax is $[f(K^F, L^F) - wL^F](1 - t^F)$. The net of tax income from capital at home is $[f(K^D, L^D) - wL^D](1 - t^D)$. The home country also levies a tax, t^H, on the income from capital earned abroad. (The rate levied on the foreign income from capital does not have to be at the same rate as the rate levied on income from capital earned at home.) This is the residency-based portion of the home country's tax on income from capital earned abroad. We assume for simplicity

2. Repatriated dividends may also be taxed by the home country under a personal income tax.

3. Recall that "commodity" here refers to both goods and factors. In general, all but one good or factor is either taxed or subsidized under optimal commodity taxation.

4. We make no distinction here between incorporated and unincorporated firms to focus on the structure of the tax on income from capital.

that capital income earned abroad is immediately repatriated and taxed.

The standard normative assumption is that the home country would want to maximize the income from capital earned by its own citizens. This is equivalent to maximizing the sum of the income net of foreign tax earned abroad plus the income earned at home. In thinking about setting its tax rates, the government understands that the foreign tax rate is a given from its perspective. If the home government levied no tax at all, its firms would earn $f[(K^F, L^F) - wL^F](1 - t^F)$ on capital invested abroad and $[f(K^D, L^D) - wL^D]$ on capital invested at home. The firms would allocate their given amounts of capital at home and abroad to maximize the sum of their earnings. This is accomplished by equalizing the marginal product of capital net of tax abroad and the marginal product of capital at home, i.e., $f'_k(K^F, L^F)(1 - t^F) = f'_k(K^D, L^D)$. Alternatively, $f'_k(K^D, L^D)/f'_k(K^F, L^F) = (1 - t^F)$.

The home government would want to maintain this condition in levying its own taxes on income from capital earned abroad and capital earned at home. With the tax on income earned abroad, t^H, and the tax on income earned at home, t^D, the firms earn $[f(K^F, L^F) - wL^F](1 - t^F - t^H)$ on capital placed abroad and $[f(K^D, L^D) - wL^D](1 - t^D)$ on capital placed at home. Under this tax regime, the firms allocate capital abroad and at home such that the marginal product of capital net of tax is equal in both countries. Therefore $f'_k(K^F, L^F)(1 - t^F - t^H) = f'_k(K^D, L^D)(1 - t^D)$ or $f'_k(K^D, L^D)/f'_k(K^F, L^F) = (1 - t^F - t^H)/(1 - t^D)$. The combined income from capital net of tax, given the foreign tax, is maximized if the home government sets $t^H = t^D(1 - t^F)$, such that $(1 - t^F - t^H) = (1 - t^F)(1 - t^D)$ and $f'_k(K^D, L^D)/f'_k(K^F, L^F) = (1 - t^F)(1 - t^D)/(1 - t^D) = (1 - t^F)$. This is equivalent to levying the home tax on the income from capital abroad net of the foreign tax, in other words allowing a deduction from income of the foreign tax paid in computing the home tax on the income abroad. The home government should treat the foreign tax liability as just another cost of doing business abroad.

This is not what governments do, however. As noted above, the typical practice of the home countries is to allow a credit against their own tax for foreign taxes. This turns out to be consistent with countries taking a worldwide perspective in which they attempt to maximize worldwide income on capital before tax rather than maximizing the incomes earned by their own firms. To see this, note that the sum of income on capital earned worldwide before tax in our simple model is $f[(K^F, L^F) - wL^F] + [f(K^D, L^D) - wL^D]$. Firms maximize this sum by allocating capital to equalize the marginal products of capital abroad and at home, i.e., $f'_k(K^F, L^F) = f'_k(K^D, L^D)$.

Return again to the framework in which the host country levies a tax t^F on the income from capital earned on the capital placed there, and the home government levies a tax t^H on the income from capital earned abroad and t^D on the income from capital earned at home. As before, the firms allocate capital to equalize the net-of-tax returns: $f'_k(K^F, L^F)(1 - t^F - t^H) = f'_k(K^D, L^D)(1 - t^D)$. Setting $t^H = t^D - t^F$ generates $f'_k(K^F, L^F) = f'_k(K^D, L^D)$, as required to maximize worldwide returns to capital.[5] This is a credit system, since it subtracts the taxes paid abroad in determining the tax liability on income from the capital earned abroad. It also gives the host country an incentive to raise its tax rate on foreign income at least to the home country's rate.[6]

The attempt to maximize income from capital worldwide may be an admirable objective, but it is undoubtedly not the conscious objective of any national government. Moreover, as noted above, it is inconsistent with the typical normative assumption that countries try to maximize a social welfare function that is defined only over their own citizens' utility functions. It also happens to be inconsistent with the literature on the international effects of corporate income taxation that followed the seminal papers by Zodrow and Mieszkowski, and Wilson.

The baseline of the ZMW model assumes that all individuals within a country are identical—they have the same tastes and own the same amount of capital and whatever other resources are in the model. Therefore, they assume that the goal of the home government is to maximize the utility of the representative consumer, the standard normative assumption with identical consumers. This is as far as they push the normative perspective, however. They simply assume that each country levies a source-based corporate income tax and only a source-based tax—there is no attempt to collect taxes on the capital income of their citizens earned in other countries. The countries then in effect play a game with each other resulting in the Nash solution: Each country sets its tax rate to maximize the utility of its own representative citizen while taking as given the tax rates set by the other countries. The primary purpose of the analysis is positive, to analyze the international economic effects of this Nash game, particularly the effects resulting from the flow of mobile capital across countries in response to their tax rates. There is no attempt in the baseline model to design a more efficient tax system.

5. This assumes $t^D > t^F$. If $t^D < t^F$, and the firms get a tax credit on their foreign income from capital only in the amount t^D (the home government does not subsidize foreign income, the standard practice), then the firms pay some tax to the host government and firms no longer allocate capital between the host and home countries to maximize worldwide capital income before tax.

6. The difference between deductions and credits in terms of what the home government is trying to maximize was first described by Peggy Richmond (1963). The analysis here is based on Gordon and Hines (2002), *op. cit.*, pp. 1943–1945.

The ZMW Model

We begin with the simplest version of the ZMW model, the one used by Zodrow and Mieszkowski in their 1986 article. The components of the model are as follows.

Production—Each of a large number (N) of countries uses capital (K) and labor (L) to produce two goods, a consumption good C and a government-provided good G. Each country produces the output with the same production function, $f(K, L)$, that exhibits constant returns to scale (CRS). A unit of output can be turned into either a unit of C or a unit of G. Because production is CRS, the production function can be written as $f(k)$, where $k = K/L$, the capital–labor ratio. Also, $f' > 0$ and $f' < 0$. Similarly, the outputs C and G are written as c and g, the amount of consumption and the government good per unit of labor.

Resources—The labor supply in each country is a constant, $\overline{L}$, which ZM set equal to one unit. The worldwide supply of capital is also constant, $\overline{K}$, but capital is completely mobile across countries. The amount of the fixed worldwide supply of capital owned by the representative consumer in the country is $\overline{k}_i$, with $\sum_{i=1}^{N}\overline{k}_i = \overline{K}$.

Utility—The representative consumer has utility function $U = U(c, g)$, which implies that G is a publicly provided private good.

Prices—All markets are assumed to be perfectly competitive. The price of the output is one. The supply price of capital on the worldwide capital market is ρ, the return that the owners of capital require on the last, marginal unit of capital. In the simplest ZMW model, each country is too small to affect ρ—it is taken as given. The government levies a source-based unit tax, t, on the use of capital in production.[7] Therefore, the user price of capital is the gross of tax price, $\rho + t$. That is, the demand for capital is a function of $\rho + t$. Because the supply price of capital is fixed at ρ, the user price of capital rises by the full amount of the tax, as illustrated in Fig. 29.1.

Market clearance—Market clearance holds implicitly within each country since there is only one consumer. Therefore, the only specific market clearance equation relates to the international allocation of the fixed supply of capital: $\sum_{i=1}^{N}k_i = \overline{K} = \sum_{i=1}^{N}\overline{k}_i$. k_i refers to the demand for capital in country i.

Government budget constraint—The government's budget constraint is $tk = g$.

The model has two main elements and they are interrelated: (1) the setting of the tax rate t within each country

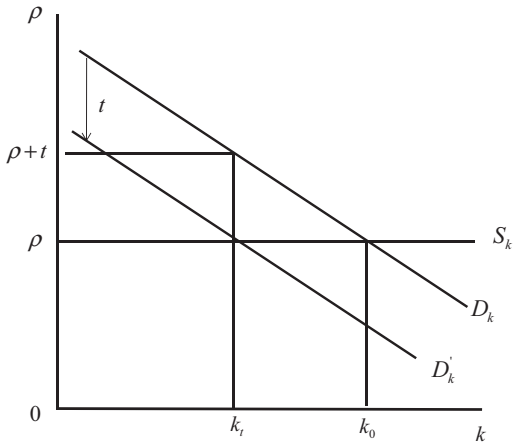

FIGURE 29.1

to maximize the utility of the representative consumer and (2) the international allocation of the fixed capital stock, which is beyond the control of any one country. Consider first the international allocation of capital.

International allocation of capital—Profit maximizing firms in each country i, faced with a user price of capital $\rho + t_i$, hire capital such that the marginal product of capital equals the user price (cost),

$$f'(k_i) = \rho + t_i \quad i = 1, \ldots N \tag{29.1}$$

These relationships, combined with the international market clearance equation for capital

$$\sum_{i=1}^{N} k_i = \overline{K} \tag{29.2}$$

solve for ρ and k_i as a function of the vector of tax rates $\overrightarrow{t} = (t_1 \ldots t_i \ldots t_N)$ levied in each country:

$$\rho = \rho(\overrightarrow{t}) \tag{29.3}$$

and

$$k_i = k_i(p(\overrightarrow{t}) + \overrightarrow{t}) = k_i(\overrightarrow{t}), \quad i = 1, \ldots N \tag{29.4}$$

Equation (29.4) determines the own and cross-price derivatives of capital with respect to tax rates within each country: $\frac{\partial k_i}{\partial t_i}$ and $\frac{\partial k_i}{\partial t_j}$, $j \neq i$.[8]

7. Although capital income taxes are ad valorem taxes, the ZMW literature usually assumes that the taxes are per-unit taxes and we follow that convention. The form of the tax makes no essential difference to the analysis.

8. $\frac{\partial k_i}{\partial t_j}$ can be derived in terms of the production functions by totally differentiating the system of profit-maximizing equations, $f_i'(k_i) - t_i = f_N'(\overline{K} - \sum_{j=1}^{N-1}k_j - t_N,\ i = 1, \ldots, N-1$. The solution, for $N \geq 3$ is $\frac{\partial k_i}{\partial t_i} = f_i''\left(1 - \frac{f_i''}{\sum_{j=1}^{N} f_j''}\right) < 0$ and $\frac{\partial k_i}{\partial t_j} = \frac{-f_i''f_j''}{\sum_{j=1}^{N} f_j''} > 0$. See Keen and Konrad (2012), *op. cit.*, pp. 7–8.

Utility Maximization—Small Country Assumption with ρ Constant

The goal of the government of country i is to

$$\begin{aligned} &\text{Max } u(c_i, g_i) \\ &(t_i) \\ &s.t\ \ t_i\, k_i = g_i \end{aligned}$$

$c = f(k_i) - (\rho + t_i)k_i + \rho\overline{k}_i$. The term $f(k_i) - (\rho + t_i)k_i$ is the return to the fixed factor, labor, equal to the value of output minus the payments to the owners of the capital at the user cost $\rho + t$. $\rho\overline{k}_i$ is the return to the capital owned by the representative consumer no matter in which country the capital is employed. Substituting for c and g into the utility function, the government's problem is

$$\begin{aligned} &\text{Max } U\big(f(k_i) - (\rho + t_i)k_i + \rho\overline{k}_i, t_i k_i\big) \\ &(t_i) \end{aligned}$$

A useful point to note before deriving the first-order condition is that the model can be expressed alternatively as a model of international trade. To see this, add $g_i = t_i k_i$ to both sides of the expression for consumption.

$$\begin{aligned} c_i + g_i &= f\big(k_i\big) - \big(\rho + t_i\big)k_i + \rho\overline{k}_i + t_i k_i \\ &= f\big(k_i\big) - \rho k_i + \rho\overline{k}_i. \end{aligned}$$

Rearranging terms,

$$\rho\big(k_i - \overline{k}_i\big) = f\big(k_i\big) - \big(c_i + g_i\big) \tag{29.5}$$

The LHS is the value of imports of capital, which is equal to the value of the exports of the consumer and publicly provided goods on the RHS (with $p_C = p_G = 1$).

Returning to the maximization of utility, each country is assumed to set its tax rate under the Nash assumption that the tax rates in all other countries remain constant. Under this assumption, and recalling that $f'(k_i) = \rho + t_i$ for profit maximization, the FOC for utility maximization is

$$U_c\left(-f''_{k_i}\frac{\partial k_i}{\partial t_i}k_i\right) + U_g\left(k_i + t_i\frac{\partial k_i}{\partial t_i}\right) = 0. \tag{29.6}$$

Rearranging terms,

$$\frac{U_g}{U_c} = \frac{f''_{k_i}\frac{\partial k_i}{\partial t_i}k_i}{k_i + t_i\frac{\partial k_i}{\partial t_i}} \tag{29.7}$$

Differentiating $f'(k_i) = \rho + t_i$, the first-order condition for profit maximization, with respect to t_i yields $f''_{k_i}\frac{\partial k_i}{\partial t_i} = 1$. Therefore,

$$\frac{U_g}{U_c} = \frac{k_i}{k_i + t_i\frac{\partial k_i}{\partial t_i}} = \frac{1}{1 + \frac{t_i}{k_i}\frac{\partial k_i}{\partial t_i}},$$

or

$$\frac{U_g}{U_c} = \frac{1}{1 + E_{k_i,t_i}} > 1, \tag{29.8}$$

where E_{k_i,t_i} is the elasticity of the demand for capital with respect to the tax rate, and assuming $0 > E_{k_i,t_i} > -1$. $\frac{U_g}{U_c}$ is the $\text{MRS}_{g,c}$ or, alternatively, $\frac{-\text{d}c_i}{\text{d}g_i}$, the marginal cost of g in terms of foregone c to the representative consumer. Since output can become either c or g, the $\text{MRT}_{g,c} = 1$; alternatively, the marginal cost of producing g is one, with $p_Q = MC_Q = 1$. Therefore, $\frac{-\text{d}c_i}{\text{d}g_i} = \text{MRS}_{g,c} > \text{MRT}_{g,c} = \text{MC}_g$; g is underprovided. This is one of the main results of the ZMW model.

The symmetric case—The literature often considers the symmetric case of all countries identical in every respect, since it leads to definitive results. In this simplest version of the model, it is used to demonstrate that all countries would be better off with an equal marginal increase in their tax rates, dt. To see this, note that under symmetry, $\overline{k}_i = \frac{\overline{K}}{N}$ and $k_i = \overline{k}_i$, for all i, in equilibrium. There can be no imports and exports of capital and thus no imports or exports of goods: $c + g = f(k)$. Equilibrium tax rates are the same worldwide, and an equal change in tax rates cannot change the amount of capital in each country. Therefore, from the profit-maximizing condition, $f'(k_i) = \rho + t_i$ and $-\text{d}\rho = \text{d}t$. The supply price of capital falls by the full amount of the common tax increase. Under these conditions, the change in utility from an equal marginal increase in tax rates is

$$\text{d}U = U_c\big(-\text{d}\rho\overline{k}_i\big) + U_g\big(k_i\text{d}t\big) \tag{29.9}$$

But $-\text{d}\rho\overline{k}_i = k_i\text{d}t$ or $-\text{d}c = \text{d}g$. With $U_g > U_c$, $\text{d}U > 0$. The allocation under the Nash equilibrium is not pareto optimal. Under symmetry, this result goes hand in hand with the result that g is underprovided within each country; the common tax rate is too low.

Fiscal Externalities

The outcome that tax rates, and hence public good provision, are too low under the Nash equilibrium is an externality problem. Each country knows that raising its tax rate, other tax rates constant, will drive capital out of the country and thus they are reluctant to raise the rate. But this ignores the benefits that its higher tax rate would confer on other countries as they receive some of the capital, benefits that are referred to in the literature as fiscal externalities. An omniscient world planner would take the externalities into account, and all tax rates would increase.

There are three possible kinds of externalities, a tax base effect, an output effect, and a terms-of-trade effect. To analyze all the three, we need to modify the simplest ZMW model above in two ways. One is to drop the assumption of symmetry, such that countries can be importers and exporters of capital (and goods). The other is to assume that each country is large enough to affect the worldwide supply price of capital, ρ.

In this expanded model, consider the effect of an increase in country j's tax rate on country i.

One external effect is on the tax base of country i. $g = t_i k_i$. $\frac{\partial g_i}{\partial t_j} = t_i \frac{\partial k_i}{\partial t_j} > 0$. The provision of g rises and this is welfare improving in and of itself.

A second external effect is that the increase in capital in country i increases its overall output through the production function $f(k_i)$. This permits the increase in g and it may also allow c to rise as well.

Whether c rises or falls depends on the third externality, the terms-of-trade effect. To see this, recall that $c_i = f(k_i) - (\rho + t_i)k_i + \rho \overline{k}_i$. Therefore, $\frac{\partial c_i}{\partial t_j} = -k_i f''_{k_i} \frac{\partial k_i}{\partial t_j} + \overline{k}_i \frac{\partial \rho}{\partial t_j}$. From the profit-maximizing condition $f'(k_i) = \rho + t_i$, $\frac{\partial \rho}{\partial t_j} = f''_{k_i} \frac{\partial k_i}{\partial t_j}$. Therefore, $\frac{\partial c_i}{\partial t_j} = -k_i f''_{k_i} \frac{\partial k_i}{\partial t_j} + \overline{k}_i f''_{k_i} \frac{\partial \rho}{\partial t_j} = (k_i - \overline{k}_i)(-f''_{k_i} \frac{\partial k_i}{\partial t_j}) \cdot -f''_{k_i} \frac{\partial k_i}{\partial t_j}$ is positive. If country i is an importer, then $(k_i - \overline{k}_i)$ is also positive, c_i rises, and given that g_i rises as well, utility increases.

The effect on the supply price of capital, $\frac{\partial \rho}{\partial t_j} = f''_{k_i} \frac{\partial k_i}{\partial t_j} < 0$ is called the terms-of-trade effect. It tends to lower consumption, and utility, because of its depressing effect on the income of the capital suppliers in the country, the $\rho \overline{k}_i$ term. But the second capital effect, the increase in k_i, dominates the terms-of-trade effect for importers of capital and c_i rises. For exporters of capital, however, the terms-of-trade effect dominates and c_i falls. Utility may or may not increase. Notice that if $\rho = \overline{\rho}$, assuming a small country, then the terms-of-trade effect disappears and utility increases from an increase in t_j because of both the tax base and output externalities.

Utility Maximization—The Large Country Case with ρ Variable

Consider, next, utility maximization under the assumption that each country is large enough that its tax policies affect the supply price of capital, ρ. Begin with the asymmetric case in which the citizens of country i own $\overline{k}_i$ of the world capital supply.

The first task is to see how the ability to affect ρ changes the marginal rate of substitution between consumption and the publicly provided good. $c_i = f(k_i) - (\rho + t)k_i + \rho \overline{k}_i$. Therefore, $\frac{\partial c_i}{\partial t_i} = f' k'_i \frac{\partial(\rho + t_i)}{\partial t_i} - (\rho + t_i) k'_i \frac{\partial(\rho + t_i)}{\partial t_i} - k_i \frac{\partial(\rho + t_i)}{\partial t_i} + \overline{k}_i \frac{\partial \rho}{\partial t_i}$, where $k'_i = \frac{\partial k_i}{\partial(\rho + t_i)}$.

From the profit-maximizing condition, $f'(k_i) = \rho + t_i$. Therefore, $\frac{\partial c_i}{\partial t_i} = -k_i \frac{\partial(\rho + t_i)}{\partial t_i} + \overline{k}_i \frac{\partial \rho}{\partial t_i}$ or $\frac{\partial c_i}{\partial t_i} = \frac{\partial \rho}{\partial t_i}(\overline{k}_i - k_i) - k_i$.

$$\frac{\partial g_i}{\partial t_i} = k_i + t_i k'_i \frac{\partial(\rho + t_i)}{\partial t_i}.$$

Therefore, the MRS_{g_i, c_i} is

$$\frac{U_{g_i}}{U_{c_i}} = -\frac{\mathrm{d}c_i}{\mathrm{d}g_i} = \frac{-\frac{\partial \rho}{\partial t_i}(\overline{k}_i - k_i) + k_i}{k_i + t_i k'_i \frac{\partial(\rho + t_i)}{\partial t_i}} \tag{29.10}$$

Dividing numerator and denominator by k, and expressing the second term in the denominator in elasticity form yields

$$\frac{U_{g_i}}{U_{c_i}} = -\frac{\mathrm{d}c_i}{\mathrm{d}g_i} = \frac{-\frac{\partial \rho}{\partial t_i}\left(\frac{\overline{k}_i}{k_i} - 1\right) + 1}{1 + \frac{t_i}{(\rho + t_i)} E_{k_i, (\rho + t_i)} \left(\frac{\partial \rho}{\partial t_i} + 1\right)} \tag{29.11}$$

The expression for the MRS_{g_i, c_i} is much more complex than Eqn (29.8), when ρ is constant.

Some further insight regarding the inefficiency of the Nash game in tax rates can be obtained from considering the symmetric case in which $\overline{k}_i = \frac{\overline{k}}{N} = k_i$, for all i. In addition, market clearance in the worldwide capital market implies that $\sum_{i=1}^{N} k_i = \overline{K}$. Differentiating the market clearance equation with respect to t_i yields $k'_i + \sum_{j=1}^{N} k'_i \frac{\partial \rho}{\partial t_i} = 0$. Given symmetry, $k' + N k' \frac{\partial \rho}{\partial t_i} = 0$ or $\frac{\partial \rho}{\partial t_i} = -\frac{1}{N}$.

Therefore, Eqn (29.11) becomes[9]

$$\frac{U_g}{U_c} = -\frac{\mathrm{d}c}{\mathrm{d}g} = \frac{+1}{1 + \frac{t_i\left(1 - \frac{1}{N}\right)}{\rho + t_i} E_{k_i, (\rho + t_i)}} > 1 \tag{29.12}$$

Once again, the allocation of the publicly provided good is inefficient. This can be illustrated directly for the two-country case. Given symmetry, $k_i = \overline{k}_i = \frac{\overline{K}}{2}$. Since there are no imports or exports of capital and goods, $c_i + g_i = f(\frac{\overline{K}}{2})$. This relationship defines the production–possibilities frontier for each country i, which has a slope of -1, since a unit of output can be either the consumption good or the publicly provided good. The frontier is pictured as line AB in Fig. 29.2.[10]

The consumption–possibilities frontier is given by Eqn (29.12). It cuts AB at point E, with the slope of indifference curve I^0 between c and g greater than one in absolute value. The efficient allocation is at point F, where the indifference curve I^1 is just tangent to AB with a slope of -1. Each country produces and consumes too much c and too little g.[11]

9. We are assuming that the elasticity is negative and small enough such that the denominator of Eqn (29.12) remains positive but less than one.

10. Absent symmetry, the position of the production–possibilities frontier would depend on the amount of capital used in each country, which would in turn depend on the vector of equilibrium tax rates.

11. William Hoyt was the first to show that, for the symmetric case, both tax rates and utility fall as the number of countries rises. That tax rates fall can be intuited from Eqn (29.12). If $N = 1$, the $\mathrm{MRS}_{c,g} = 1 = \mathrm{MRT}_{c,g}$, the efficient case. As N increases, more c has to be sacrificed per additional unit of g, because the external effects of increasing the tax rate rise. The countries react by lowering their tax rates and the provision of g. The loss of utility requires a deeper proof and intuition, which Hoyt provides. But it also results from the increasing externality effects as N increases and the increasing underprovision of g (Hoyt, 1991).

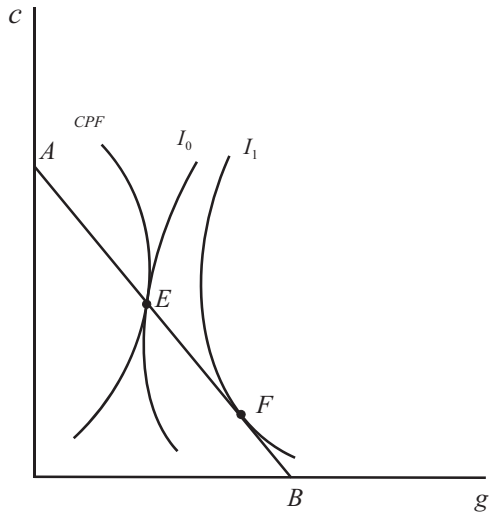

FIGURE 29.2

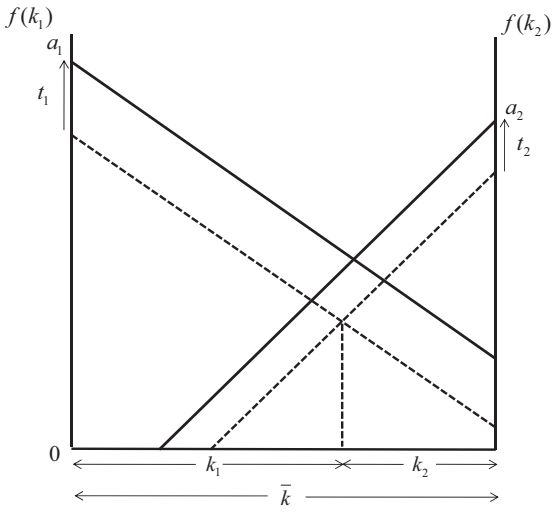

FIGURE 29.3

Reaction Functions

General results are difficult to come by in the asymmetric case. Therefore, to illustrate some possibilities, the literature often follows two simplifications proposed by David Wildasin in a 1991 paper. Wildasin assumed that each country's production function is quadratic and that utility is linear in either the consumption or the publicly provided good. Among other things, these assumptions permit solutions of closed-form reaction functions in the two-country case.[12]

First, assume that each country has a quadratic production function $f(k) = ak - 1/2\ k^2$. Therefore, $f_k' = a_i - k_i = \rho + t_i, \quad i = 1,2$ from profit maximization. Alternatively, $\rho = a_1 - k_1 - t_1 = a_2 - k_2 - t_2$. These relationships, along with the worldwide market clearance equation, $k_1 + k_2 = \overline{K}$, determine the allocation of capital for given tax rates t_1 and t_2 and ρ, as illustrated in Fig. 29.3. The figure assumes that $a_1 - k_1 > a_2 - k_2$, so that $t_1 > t_2$ and $k_1 > k_2$.

Assume that utility $U^i = c_i + G(g_i)$, for $i = 1,2$, is linear in the consumption good. To simplify further, assume $G(g_i) = (1+\lambda)g_i$, with $\lambda > 0$. As before, $c_i = f(k_i) - (\rho + t_i)\,k_i + \rho\overline{k}_i$ and $g_i = t_i k_i$. Therefore, from utility maximization

$$\frac{\partial U}{\partial t_i} = -f_i''\frac{\partial k_i}{\partial t_i}k_i + \frac{\partial \rho}{\partial t_i}\overline{k}_i + (1+\lambda)\left(k_i + t_i\frac{\partial k_i}{\partial t_i}\right) = 0, \quad i = 1,2. \tag{29.13}$$

In addition, $\rho = a_1 - k_1 - t_1 = a_2 - k_2 - t_2$. Therefore, $-\frac{\partial k_1}{\partial t_1} - 1 = -\frac{\partial k_2}{\partial t_1}$. But any change in k_1 must lead to an equal change in k_2 in the opposite direction in the two-country case with a fixed worldwide capital stock. Thus, $-\frac{\partial k_1}{\partial t_1} = \frac{\partial k_2}{\partial t_1}$ and $\frac{\partial k_1}{\partial t_1} = -\frac{1}{2}$. In addition, from profit maximization, $\frac{\partial \rho}{\partial t_1} = f_1''\frac{\partial k_1}{\partial t_1} - 1 = -1(-\frac{1}{2}) - 1 = -\frac{1}{2}$. (Similarly, $\frac{\partial k_2}{\partial t_2} = -\frac{1}{2}$ and $\frac{\partial \rho}{\partial t_2} = -\frac{1}{2}$.) Given these relationship, Eqn (29.13) becomes

$$\frac{\partial U}{\partial t_i} = -\frac{1}{2}k_i - \frac{1}{2}\overline{k}_i + (1+\lambda)k_i - \frac{1}{2}(1+\lambda)t_i = 0 \tag{29.14}$$

Next, since $a_i - k_i - t_i = \rho$, then, $(a_1 + a_2) - (k_1 + k_2) - t_1 - t_2 = 2\rho = A - \overline{K} - t_1 - t_2$, where $A = a_1 + a_2$ and $k_1 + k_2 = \overline{K}$. Therefore, $\rho = \frac{A}{2} = \frac{\overline{K}}{2} - \frac{(t_1+t_2)}{2}$. But $k_i = a_i - t_i - \rho$. Substituting for k_i in Eqn (29.14) yields

$$\begin{aligned}\frac{\partial U}{\partial t_i} = &-\frac{1}{2}\left[a_i - t_i - \frac{A}{2} + \frac{\overline{K}}{2} + \frac{(t_1+t_2)}{2}\right] \\ &-\frac{1}{2}\overline{k}_i + (1+\lambda)\left[a_i - t_i - \frac{A}{2} + \frac{\overline{K}}{2} + \frac{(t_1+t_2)}{2}\right] \\ &-\frac{1}{2}t_i(1+\lambda) = 0.\end{aligned} \tag{29.15}$$

Equation (29.15) is symmetric in t_1 and t_2. Therefore, let $i = 1$ and solve Eqn (29.15) for t_1 as a function of t_2. The solution, after considerable manipulation, is

$$t_1 = \frac{(1+2\lambda)(a_1 + \frac{\overline{K}}{2} - \frac{A}{2}) - \overline{k}_1}{\frac{3}{2} + 2\lambda} + \left(\frac{\frac{1}{2} + \lambda}{\frac{3}{2} + 2\lambda}\right)t_2. \tag{29.16}$$

Equation (29.16) is the reaction function of country 1's tax rate in terms of country 2's tax rate under the Nash assumption. It is linear, having the form $t_1 = A + Bt_2$, and it is symmetric for country 2. Equation (29.16) yields a number of interesting results.

12. Wildasin (1991). Although the assumptions are due to Wildasin, the derivation here follows that of Keen and Konrad (2012) *op. cit.*, pp. 14–21.

1. *Strategic complements*—The coefficient on t_2 in Eqn (29.16) is positive, implying that t_1 and t_2 are strategic complements: The reaction functions are upward sloping. That tax rates rise and fall together, coupled with the result that tax rates are set too low because countries ignore the positive fiscal externalities in other countries that an increase in their tax rates would generate, implies that Nash competition can lead to a "race to the bottom." A cut in one country's taxes leads to a cut in the other country's taxes. Indeed, as noted earlier, William Hoyt showed that an increase in the number of countries leads to a decrease in tax rates and utility for the case of symmetric countries (see footnote 11). The intuition is that the elasticity of capital with respect to the user cost of capital rises with the number of countries, thereby increasing the outflow or inflow of capital for any given change in tax rates. The incentive is for each country to decrease its tax rates in the Nash game.

   In fact, average statutory corporate tax rates have declined steadily and substantially since the early 1990s. From 1993 to 2012, KPMG reports that the average rate in all the countries it surveys fell from 38% to 24%; for the OECD countries, from 38% to 25%; and for the EU countries, from 38% to 23%. An interesting question is whether this downward trend will continue, to the point the countries essentially give up trying to raise tax revenues from corporations given the mobility of capital.[13]

   The reaction functions also offer another perspective on the inefficiency of tax competition. Figure 29.4 pictures the reaction functions, $t_1(t_2)$ and $t_2(t_1)$, for two symmetric countries, that is, two countries that have the same values a, λ, and $\overline{k}$.

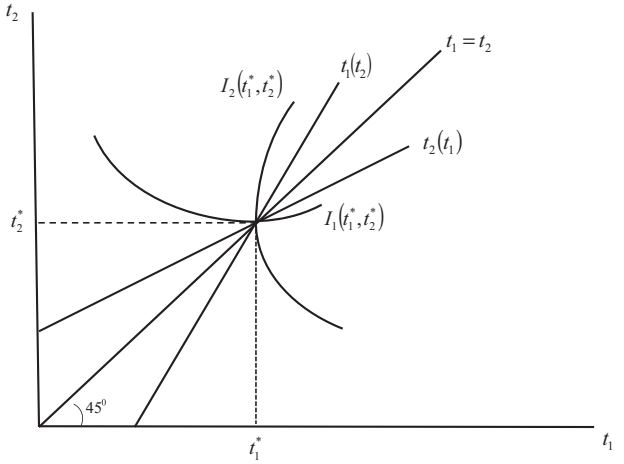

FIGURE 29.4

   The reaction functions are assumed to hit the axes above zero because the publicly provided good g is considered valuable enough that each country would levy some tax to finance g even if the tax rate in the other country were zero. The equilibrium is (t_1^*, t_2^*), at the intersection of the two reaction functions, on the 45° line because the two countries are symmetric. Note also that the slope of the indifference curve in t_1 and t_2 for country 2 must be vertical at (t_1^*, t_2^*) since a marginal change in t_2 at the optimum cannot have a first-order effect on utility. Similarly, the slope of indifference curve for country 1 must be horizontal at (t_1^*, t_2^*). The area above and between the two indifference curves defines the set of tax rates that would be pareto improving, demonstrating the inefficiency of the Nash equilibrium.

2. *Exporting and importing countries*—Suppose that the countries are identical in every respect except that citizens in country 1 own more of the fixed worldwide capital stock: $\overline{k}_1 > \overline{k}_2$. That is, at equal tax rates, with an equal amount of capital employed in each country, country 1 would be an exporter of capital and country 2 would be an importer of capital. Instead of setting equal tax rates, however, t_1 would be less than t_2. This can be seen by noting that the reaction functions for t_1 and t_2 have the form $t_1 = C + Dt_2 - E\overline{k}_1$ and $t_2 = C + Dt_1 - E\overline{k}_2$. The Nash equilibrium is only possible if $t_1 < t_2$.[14] The intuition is that raising t_1 leads to an increase in t_2, and for the relatively capital-rich, exporting country, this implies that some of the increase in taxes paid by its citizens will accrue to the other country. Therefore, the exporting country has an incentive not to be so aggressive in raising taxes. Conversely, the relatively capital-poor importing country bears less of a cost in raising its tax rates and will therefore be more aggressive in raising its rates.

   One potential difficulty with incentives under unequal tax rates is illustrated in Fig. 29.5.

   The intersection of the different reaction functions leads to t_1^* less than t_2^*. When this outcome occurs, there is often a call to equalize or "harmonize" tax rates across countries, presumably at a higher level, to avoid unfair and destructive competition for capital. For example, this has been a long-standing issue in the European Union (EU) led by high-tax countries such as Germany and France, which do not want to lose capital to low-tax countries such as Ireland. But, as the figure indicates, the 45° − line of equal tax rates lies below the region of pareto improvement in tax rates defined

13. KPMG's Corporate Tax Rate Survey (2006). The rates from 2006 to 2012 are available at www.kpmg.com/global/en/services/pages/corporate-tax-rates-table.aspx.

14. Solving for t_1 and t_2 indicates that $t_1 < t_2$ only if $\overline{k}_1 E(1-D) > \overline{k}_2 E(1-D)$, where $D = \left(\frac{\frac{1}{2}+\lambda}{\frac{3}{2}+2\lambda}\right) < 1$ and $E = \frac{1}{\frac{3}{2}+2\lambda} > 0$.

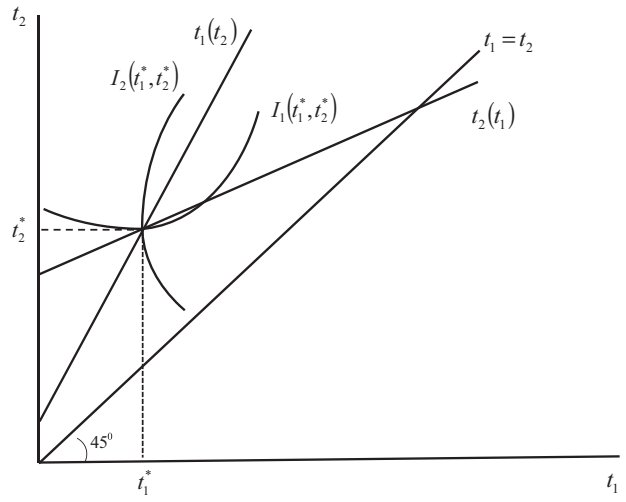

FIGURE 29.5

by the two indifference curves through the equilibrium tax rates. The low-tax country would object to a harmonization policy, as indeed the low-tax countries such as Ireland have done in the EU.

3. *Larger versus smaller economies*—Suppose that the two countries are identical in every respect except that $a_1 > a_2$—country 1 has the more productive economy and therefore attracts more capital, as demonstrated in Fig. 29.3. The term a_i appears only in the constant term in Eqn (29.16). Therefore, $a_1 > a_2$ implies that $t_1 > t_2$. The smaller countries have the lower tax rates, and vice versa. This is one of the most important results in the international capital taxation literature and it is consistent with reality.

The high-tax countries throughout the world do tend to be those who have the larger, capital-intensive economies. The intuition for this is that capital is durable, and once in place, taxing it is essentially a lump-sum tax because old capital in place cannot respond to the taxes on it. The presence of old capital leads to the so-called time-inconsistency problem, that countries want to keep taxes on capital low to attract investment, but then tax capital once it is in place up to 100% because that tax would be a lump-sum tax. Governments would not do that because investors will adjust to confiscatory taxes on old capital by refusing to invest in the country in the first place. Nonetheless, old capital is a ready source of tax revenues and explains why capital-rich countries have an incentive for high tax rates despite driving some investment away each year. If there are agglomeration economies in which the return to investment rises if the investment takes place in an environment with a lot of capital already in place—think of high-tech companies wanting to be near each other—then the high tax rate on capital may not be so discouraging to investment. It would take a multiperiod model to fully describe these incentives for higher tax rates on capital, but our simple model points in the right direction.[15]

SUMMARY

The simple models in this section were able to show, or at least to suggest, three of the main results from the ZMW model as follows:

1. It pays to be small. The smaller countries have lower tax rates, which allow them to benefit from the fiscal externalities inherent in the setting of corporate income tax rates. They attract more investment, which increases output per capita and the per capita consumption of the public good and, since they are likely to be importers of capita, consumption per capita of the consumer good as well.
2. Nash competition over tax rates leads to underprovision of the public good along with tax rates that are too low. In the symmetric case, all countries would benefit from a common marginal increase in tax rates. In the realistic asymmetric case, however, the low-tax countries might not benefit from an increase in tax rates. They are likely to resist efforts to equalize—"harmonize"—tax rates at higher values, as the larger, capital-rich countries would prefer.
3. Reaction functions defined over tax rates are likely to be upward sloping in the Nash setting. This leads to the possibility of a race to the bottom in tax rates, and statutory tax rates have fallen considerably since the 1990s.

MULTINATIONAL ENTERPRISES

The rise of MNEs leads to the natural research question: What advantages are there for firms to establish subsidiary companies and plants throughout the world? A number of answers come immediately to mind—ready access to raw materials and other resources that are not available in the home country, a desire to jump over tariff walls that exist in countries in which firms sell their products, and potential tax advantages, to name three of the more important ones.

15. The result that smaller countries have the lower taxes on capital is originally due to John Wilson (1991). His paper has a model of two countries that use both labor and capital to produce output. One country is larger in the sense that it has a larger labor force. In the context of that model, he shows that the country with the smaller population has the lower tax rate on capital and the higher utility level. Moreover Wilson did not have to place any restrictions on the utility function that the government is trying to maximize. His proof established the principle that in the matter of tax competition, it pays to be small. This was seen as a striking result because, as noted above, the basic ZMW model can be viewed as a model of international trade in goods and capital, and a standard result in the trade literature is that it pays to be big. Big countries with some monopoly power can use tariffs to manipulate the terms of trade to their advantage.

It is the potential tax advantages that have attracted the attention of public sector economists.

Differences in source-based corporation income tax rates throughout the world give MNEs incentives to choose locations to minimize their overall tax liabilities. Two incentives have received the most analytical attention because they appear to be the most relevant to firms' actual location decisions: tax avoidance by shifting taxable accounting profits across countries and the location of specialized resources that generate pure economic profits that cannot easily be competed away. We will consider only tax avoidance, since the analysis of specialized resources is quite similar to that of tax avoidance.

Four preliminary observations about the MNE tax literature are worth making to set the context of the analysis. The first is that the tax competition in the ZMW tradition is concerned with the effect of differences in source-based corporate income taxes on the real investment decisions of firms. The MNE literature, in contrast, has a large component that is strictly financial in nature, focusing on where firms report accounting profits that are subject to tax. This is the component that we will discuss.

The second observation is that the purely financial incentives related to accounting profits depend on differences across countries in their statutory tax rates, and often the average statutory rates, whereas the ZMW real investment effects depend on differences in the marginal effective tax rates. Average statutory rates and marginal effective tax rates can be quite different, with features such as graduated tax rates, complex depreciation allowances, and investment tax credits built into the corporation income tax structures.

The third observation is that the willingness of countries to share tax information with each other is a central feature of firms' location decisions. Firms have a natural incentive to locate subsidiaries in host countries that are willing to conceal some or all of the taxable profits from the tax authorities in the home country. With the profits concealed, the home countries cannot easily tax the profits again under either their corporation income taxes or their personal income taxes when the profits are repatriated as dividends to the citizens in the home country.

The final observation is that MNEs make liberal use of tax havens to shield profits from taxation. A tax haven is difficult to define precisely, but it tends to have three features: it is willing to set very low, even zero tax rates on corporate profits in return for collecting fees from MNEs that locate there; the goal of the tax haven is to attract accounting profits rather than real investment activity, which profits they can relend, much as private banks relend deposits for profit; and it is reluctant to share information on the profits from firms locating there. Tax havens have received considerable attention in the MNE tax literature, including why they exist and their benefits and costs. Our analysis of tax avoidance will take all four observations into account.

Tax Avoidance

MNEs can fairly easily shift accounting profits from high-tax to low-tax countries to reduce their overall tax liabilities. There are a number of ways to do this. One common method is to borrow to finance their investments worldwide in the high-tax countries to take advantage of the ability to deduct interest payments from taxable profits, a common feature of corporation income taxes, and to lend in low-tax countries to reduce the tax liability on their interest receipts. Another common method is through transfer pricing. Suppose one of an MNE's subsidiaries located in country A makes an intermediate product that it sends to another of its subsidiaries located in country B, where it is used to produce a final product sold in country B. The MNE has to set a price on the intermediate product for tax accounting purposes in both countries. These internally set prices are called transfer prices. The international convention regarding the setting of transfer prices is that they should be set at the prices they would command if sold in the market to an unrelated third party. In truth, however, MNEs have quite a bit of latitude in setting their transfer prices, which they can then coordinate with their location decisions. In the example above, if country A has a low tax rate and country B a high tax rate, the firm can reduce its overall tax liability by setting an artificially high price on the intermediate product. The high price increases accounting profits of the subsidiary in low-tax country A and increases the costs/reduces the accounting profits of the subsidiary in high-tax country B. We assume in what follows that MNEs can shift taxable accounting profits across countries by such methods to take advantage of differences in tax rates.

We begin with the firm's perspective and then bring in the countries' perspectives.

The Firms' Perspective

The standard model of multinational tax shifting assumes that MNEs can shift any amount of their actual profits across countries to generate a set of artificial reported taxable profits, but that there is a cost of doing so.[16] The model is similar to the model of tax evasion by individuals described in Chapter 15, which also assumes that tax avoidance is costly. Assume that an MNE has subsidiaries located in N different countries with flat rate taxes on reported profits of t_i, for $i = 1,\ldots, N$. The actual profits earned in each country are ϕ_i. Through such practices as arbitrary transfer pricing, the firm can shift ψ_i of its profits earned worldwide to ($\psi_i > 0$) or from ($\psi_i < 0$) country i, subject to the constraint that $\sum_{i=1}^{N} \psi_i = 0$. The cost of shifting

16. The analysis in this section is taken from Gordon and Hines (2002), *op. cit.*, pp. 1979–1982.

profits ψ_i is $\gamma(\frac{\psi_i}{\phi_i})\psi_i = \gamma\frac{\psi_i^2}{\phi_i}$, with $\gamma > 0$. That is, the cost depends on both the amount of profit shifted and the proportion of shifted profits to the true profits earned in country i. It is also assumed to be deductible in computing taxable profits. The reported accounting profit subject to tax in country i is $\pi_i = \phi_i + \psi_i - \gamma\frac{\psi_i^2}{\phi_i}$. The MNE's goal is to maximize worldwide after-tax profits by shifting taxes subject to the constraint that the amount of profits shifted adds to zero.

Setting up the Lagrangian equation,

$$\text{Max } \pi_i = \sum_{i=1}^{N}(1-t_i)\left(\phi_i + \psi_i - \gamma\frac{\psi_i^2}{\phi_i}\right) - \lambda\left(\sum_{i=1}^{N}\psi_i\right)$$

(ψ_i)

The FOC are $(1-t_i)(1-2\gamma\frac{\psi_i}{\phi_i}) - \lambda = 0$. Solving for ψ_i,

$$\psi_i = \frac{(1-t_i-\lambda)}{2\gamma(1-t_i)}\phi_i \tag{29.17}$$

λ represents the value of the average after-tax profit rate across the countries. Therefore, ψ_i is positive if $t_i < 1-\lambda$, the average tax rate. The intuitive result is that MNEs shift reported taxes to relatively low-tax countries. This is supported by the empirical literature, which consistently finds that observed profit rates per unit of capital tend to be low in high-tax countries and high in low-tax countries.

A less-comforting implication of the model from an empirical perspective is the effect of tax shifting on real investment activity. Suppose the worldwide opportunity cost of capital is ρ, as in the ZMW model in the first section of the chapter, and that MNEs are each too small to affect it. If, as in the first section, f_k is the marginal product of capital, then the firm invests until $f_k'(1-t_i) = \rho$ or $f_k' = \frac{\rho}{(1-t_i)}$. f_k' is the source of the true profits. But reported profits $\pi_i = \left(\phi_i + \psi_i - \gamma\frac{\psi_i^2}{\phi_i}\right)$ also depend on the true profits. Therefore, an increase in capital affects reported profits by $\frac{d\phi_i}{dk_i}\frac{d\pi_i}{d\phi_i}$, f_k' times $\frac{d\pi_i}{d\phi_i} = \left(1+\frac{\gamma\psi_i^2}{\phi_i^2}\right)$. The new equilibrium investment condition for the MNE is $f_k'(1-t_1)(1+\frac{\gamma\psi_i^2}{\phi_i^2}) = \rho$ or $f_k' = \frac{\rho}{(1-t_1)\left(1+\frac{\gamma\psi_i^2}{\phi_i^2}\right)}$ For local firms either in host or home countries, which cannot take advantage of tax shifting, the user cost of capital is $\frac{\rho}{(1-t_i)}$ Therefore, the MNE has a lower user cost of capital than the local firms if $\psi_i \neq 0$, that is, than local firms in countries with both lower- and higher-than-average tax rates. The advantage they have in the high-tax countries is that they can shift some of the true profits to low-tax countries, which, simultaneously, is their advantage of being able to set up subsidiaries in the low-tax countries. This implies that FDI should be attracted to both low- and high-tax countries, whereas the empirical literature has consistently found that only relatively low-tax rates attract FDI, other things being equal.[17]

Strategic Considerations—The Countries' Perspective[18]

So far we have been considering the location decision from the point of view of the MNEs. But the countries are also players in this game. They adjust their tax rates based on their perception of how the MNEs will shift profits in reaction to their tax rates. Assume that there are only two countries, 1 and 2, and that an MNE earns (true) profits in both countries, ϕ_1 in country 1 and ϕ_2 in country 2. Suppose that $t_1 < t_2$ and, for simplicity, that the owners of the MNE can avoid taxes on repatriated profits. Therefore, the MNEs have an incentive to shift a portion of their profits from country 2 to country 1 for tax purposes.

Let S be the proportion of profits shifted. As above, assume that the cost of shifting profits depends on the proportion of true profits shifted and the amount of profits shifted, with the cost parameter $\gamma = {}^1\!/_2\ \delta$, for simplicity. Given the definition of S, the cost is ${}^1\!/_2\ \delta S^2\phi_2$.[19] Here, though, the costs of shifting profits are not tax deductible, again just to simplify the calculations.

The MNE's goal is to set S to maximize worldwide after-tax profits:

$$\text{Max } \phi = \phi_1 + \phi_2 - t_1(\phi_1 + S\phi_2) - t_2(\phi_2 - S\phi_2) - 1/2\delta S^2\phi_2$$

(S)

Differentiating, $\frac{d\phi}{dS} = -t_1\phi_2 + t_2\phi_2 - \delta S\phi_2 = 0$. Therefore,

$$S = \frac{t_2 - t_1}{\delta} \tag{29.18}$$

17. The analysis ignores the issue that profits of the MNEs may be taxed again when they are repatriated to the home country, either under a corporation or a personal income tax. One way to avoid this is for the MNE to set up the parent firm in a zero-tax haven that has no tax on any form of income. In fact, profits are usually not taxed by the home country until they are repatriated, and firms can use a number of techniques to considerably delay repatriation. One obvious method is to reinvest some or all of the profits earned in the subsidiary located in the host country.

18. The analysis in this section follows along the lines suggested by Keen and Konrad (2012) *op. cit.*, pp.46–49. It is a variation of their model of sales tax competition.

19. In terms of the previous model, $S = \frac{\psi_2}{\phi_2}$. Therefore, $S^2\phi_2 = \frac{\psi_2^2}{\phi_2^2}\phi_2 = \frac{\psi_2^2}{\phi_2}$, as above.

Consider, next, the strategic interaction of the two countries. Assume that they know the profit-shifting rule Eqn (29.18) of the MNE and that their objective is to maximize their individual tax revenues. They play a Nash game with each other in the ZMW tradition, setting their tax rates to maximize their revenues on the assumption that the tax rate of the other country is fixed. The first-order conditions of the tax revenue maximization define their reaction functions in terms of the two tax rates.

Country 1's tax revenues, given S, are $T_1 = t_1(\phi_1 + (\frac{t_2-t_1}{\delta})\phi_2)$. Maximizing T_1 with respect to t_1 yields $\frac{dT_1}{dt_1} = \phi_1 + (\frac{t_2-t_1}{\delta})\phi_2 - t_1\phi_2(\frac{1}{\delta}) = 0$. Solving for t_1 generates the reaction function for country 1 (R_1):

$$R_1 : t_1 = \frac{1}{2}\delta\frac{\phi_1}{\phi_2} + \frac{1}{2}t_2 = \frac{1}{2}\delta\theta + \frac{1}{2}t_2, \quad \text{with } \theta = \frac{\phi_1}{\phi_2}.$$

Similarly, country 2's tax revenues are $T_2 = t_2\phi_2(1 - \frac{t_2-t_1}{\delta})$. Maximizing T_2 with respect to t_2 yields $\frac{dT_2}{dt_2} = \phi_2(1 - \frac{t_2-t_1}{\delta}) - t_2\phi_2(\frac{1}{\delta}) = 0$. Solving for t_2 generates the reaction function for country 2 (R_2):

$$R_2 : t_2 = \frac{1}{2}(\delta + t_1)$$

Solving R_1 and R_2 for t_1 and t_2 yields

$$t_1 = \delta\left(\frac{2}{3}\theta + \frac{1}{3}\right) \tag{29.19}$$

and

$$t_2 = \delta\left(\frac{1}{3}\theta + \frac{2}{3}\right) \tag{29.20}$$

We have assumed that $t_1 < t_2$, such that the MNE shifts reported profits to country 1. From Eqns (29.19) and (29.20), $t_1 < t_2$ only if $\delta\left(\frac{2}{3}\theta + \frac{1}{3}\right) < \delta\left(\frac{1}{3}\theta + \frac{2}{3}\right)$ or $\theta = \frac{\phi_1}{\phi_2} < 1$. This echoes the result of the ZMW model that the smaller country has the lower tax rates in a Nash game played over tax rates. It therefore lends support to the existence of tax havens.

Concealing Information[20]

The willingness of host countries to conceal some or all information on earnings from home countries is an obvious source of attraction to MNEs. Departments of revenue in the home country have very limited means of discovering foreign source income if the foreign tax authorities will not cooperate with them. Countries form bilateral tax agreements that encourage each other to reveal taxable income, but these agreements are not entirely successful. In particular, tax haven countries are reluctant to participate in such agreements.

20. The analysis in this section follows closely the presentation in Keen and Konrad (2012) *op. cit.*, pp. 56–59.

The standard model of income concealment is similar in structure to the model of tax avoidance above. There is one additional layer of complexity to the standard Nash game, however: At least one of the countries has to decide how much income to conceal along with the choice of its tax rate. As with tax avoidance, we consider a highly simplified two-country model that reveals the main issues involved with concealing income.

Assume that there are two countries, 1 and 2, with country1 the low-tax country: $t_1 < t_2$. Firms (or households) in high-tax country 2 have a given amount of saving, S, that they can place in either country.[21] s_1 is placed in country 1, s_2 in country 2, with $S = s_1 + s_2$. Saving placed in country 1 is taxed at rate t_1. Saving remaining in country 2 is taxed at rate t_2. The tax authorities in country 2 also want to tax s_1, which they would do at rate $(t_2 - t_1)$, giving a credit for the tax paid in country 1. But the tax authorities in country 1 reveal only a portion λ of s_1. Therefore, the tax paid and collected by country 2 on s_1 in country 1 is $\lambda s_1(t_2 - t_1)$. There is also a cost of placing saving in country 1 equal to $C(s_1)$, with C', $C'' > 0$. $C(s_1)$ is not tax deductible. The objective of the firm is to choose s_1 to minimize the sum of the total tax revenues paid to both countries and the costs incurred in shifting some saving to country 1.

$$\underset{(s_1)}{\text{Min}}\, T = t_2(S - s_1) + t_1 s_1 + \lambda(t_2 - t_1)s_1 + C(s_1)$$

Alternatively, the firm chooses s_1 such that the savings in tax revenue from shifting some saving to country 1 minus the cost of shifting the income is maximized. The firm would pay a tax of $t_2 s_1$ if s_1 were kept at home and a tax of $[t_1 + \lambda(t_2 - t_1)]s_1$ on s_1 if placed in country 1. The tax savings is $s_1[(t_2 - t_1) - \lambda(t_2 - t_1)] = (1 - \lambda)(t_2 - t_1)s_1$. Therefore, the firm chooses s_1 such that $(1 - \lambda)(t_2 - t_1)s_1 - C(s_1)$ is maximized.

The FOC for this problem is $(1 - l)(t_2 - t_1) = C'(s_1)$.

Assume the same quadratic cost function from the tax avoidance model, $C(s_1) = \frac{1}{2}\delta s_1^2$. Therefore, $(1 - \lambda)(t_2 - t_1) = \delta s_1$, or

$$s_1 = \frac{(1 - \lambda)(t_2 - t_1)}{\delta}. \tag{29.21}$$

Note, for future reference, the three partial derivatives of s_1:

$\frac{\partial s_1}{\partial \lambda} = -\frac{1}{\delta}(t_2 - t_1) < 0$; the more country 1 reveals information about income, the less attractive it becomes for the firm to place its saving there.

21. S is defined such that a unit of S yields one unit of taxable income. This assumption avoids including an interest rate in the analysis that has no essential role to play. The further assumption that saving takes place in only one country avoids the complication of savings flows moving in both directions.

$\frac{\partial s_1}{\partial t_1} = -\frac{1}{\delta}(1-\lambda)<0$; the higher is t_1, the less attractive it becomes for the firm to place its saving there.

$\frac{\partial s_1}{\partial t_2} = \frac{1}{\delta}(1-\lambda) > 0$; the higher is t_2, the more attractive it becomes for the firm to place its saving in country 1.

As with the tax avoidance model, this is the first stage of the Nash game. The firm makes its saving decision given the two tax rates and the proportion of the income earned in country 1 that will be revealed to the tax authorities in country 2. In the next stage, the two countries are assumed to know the firm's saving decision Eqn (29.21) and will use that information to maximize their tax revenues. In accordance with the Nash framework, country 2 chooses t_2 to maximize its tax revenue assuming that both t_1 and λ are given. That maximization yields its reaction function $t_2 = t_2(t_1, \lambda)$. The decision process of country 1 is less certain. On the one hand, knowing Eqn (29.21), it could set λ exogenously in a second stage and then choose t_1 to maximize its tax revenues, given t_2 and λ. On the other hand, it could choose t_1 and λ simultaneously to maximize its tax revenues, given t_2. The first process is simpler and will suit our purposes.

The tax revenues for the two countries are

$$T_1 = t_1 s_1 = = t_1 \frac{(1-\lambda)(t_2 - t_2)}{\delta}$$

$$T_2 = t_2 s_2 + \lambda(t_2 - t_1)s_1.$$

Substituting $S - s_1$ for s_2, and rearranging terms

$$T_2 = t_2 S - \Big[\lambda t_1 + (1-\lambda)t_2\Big]s_1$$
$$= t_2 S - [\lambda t_1 + (1-\lambda)t_2]\frac{(1-\lambda)(t_2 - t_1)}{\delta}.$$

Suppose country 1 undertakes a marginal increase in λ, perhaps succumbing to pressure from the international community to reveal more tax information. Moreover, it holds its tax rate constant despite knowing that it will receive less saving from country 2 and therefore that its tax revenues will decrease. That is, it moves off its reaction function. This simple case is useful for considering some possibilities. We continue to assume that country 2 remains on its reaction function.

From above, $T_2 = t_2 S - [\lambda t_1 + (1-\lambda)t_2]s_1$. Given that $t_1$is held constant, and that marginal changes in t_2 cannot affect T_2 if country 2 is on its revenue-maximizing reaction function, $\frac{dT_2}{d\lambda} = -[\lambda t_1 + (1-\lambda)t_2]\frac{\partial s_1}{\partial \lambda} + s_1(-)[t_1 - t_2] = -[\lambda t_1 + (1-\lambda)t_2]\frac{\partial s_1}{\partial \lambda} + s_1[t_2 - t_1]$. With, $\frac{\partial s_1}{\partial \lambda} < 0$, $\frac{dT_2}{d\lambda} > 0$. Tax revenue rises in country 2 as less saving is diverted to country 1.

The effect of the exogenous increase in λ is not as straightforward on the tax revenue for country 1, where $T_1 = t_1 s_1$. The increase in λ lowers its tax revenue by driving some of country 2′s saving away. But, with country 2 on its reaction function, a change in λ will lead to a change in t_2. Therefore, $\frac{dT_1}{d\lambda} = t_1\left[\frac{\partial s_1}{\partial \lambda} + \frac{\partial s_1}{\partial t_2}\frac{dt_2}{d\lambda}\right]$. The first term inside the brackets is negative. But $\frac{\partial s_1}{\partial t_2}$ is positive. Therefore, if $\frac{\partial t_2}{\partial \lambda}$ is positive, the entire term in brackets could be positive, in which case the exogenous increase in λ increases tax revenue in country 1. Moreover, $\frac{\partial t_2}{\partial \lambda}$ could well be positive. With country 1 now less attractive to saving, country 2 could raise its tax rates to increase its revenue even more than would result from just an increase in λ. But raising its tax rate does drive some saving to country 1. If the diversion is large enough, it can counteract the diversion in the opposite direction caused by the increase in λ.

This is in fact what happens in our simple model with a quadratic cost function for reasonable values of the parameters. To sketch out the result:

$$T_2 = t_2 S - [\lambda t_1 + (1-\lambda)t_2]\frac{(1-\lambda)(t_2 - t_1)}{\delta}$$

Given λ and t_1, country 2 maximizes tax revenue by setting $\frac{\partial T_2}{\partial t_2} = 0 = S - \frac{(1-\lambda)}{\delta}\Big[(\lambda t_1 + (1-\lambda)t_2) + (1-\lambda)(t_2 - t_1)\Big]$. Solving for t_2 yields

$$t_2 = \frac{\delta S}{(1-\lambda)(2-2\lambda)} - \frac{(2\lambda - 1)}{(2-2\lambda)}t_1, \qquad (29.22)$$

country 2′s reaction function in terms of λ and t_1. Notice that $\lambda < 1/2$ is required for the reaction function to be upward sloping, a reasonable condition. With t_1 constant, it can be shown that (after much manipulation)

$$\frac{\partial t_2}{\partial \lambda} = \frac{-2\delta S(2\lambda - 2)}{((1-\lambda)(2-2\lambda))^2} - \frac{2}{(2-2\lambda)^2}t_1 \qquad (29.23)$$

This derivative is almost certainly positive assuming $\lambda < 1/2$ and large enough such that $\frac{\partial s_1}{\partial t_2}\frac{dt_2}{d\lambda} = \frac{1}{\delta}(1-\lambda)\frac{dt_2}{d\lambda} > \frac{\partial s_1}{\partial t_2} = -\frac{1}{\delta}(t_2 - t_1)$ in absolute value, making $\frac{dT_1}{d\lambda}$ positive.

This result is almost surely due to the simplicity of the model and the assumption of only two countries. The expected result is that the diversion of saving back to country 2 by raising λ would exceed the diversion in the other direction by an induced increase in t_2, and therefore that T_1 would fall. This is especially so if there are a number of tax havens that savers in country 2 could choose from, as indeed there are; they would simply shift s_1 from country 1 to other low-tax countries. If this expectation is correct, then it suggests that the small low-tax tax havens are likely to resist the calls from the high-tax countries to reveal the

incomes of the MNEs reported in the tax havens, which has been the case.

CONCLUDING OBSERVATIONS

The EU and the OECD have long-expressed concerns about tax competition, which they view as harmful. In 2007, the European Commission established a nonbinding "Code on Conduct on Business Taxation" that tried to prevent preferential treatment of foreign investment and nontransparent tax practices, but it took no position on equalizing—"harmonizing"—corporate tax rates to avoid competition over tax rates. In 1998, the OECD published a report entitled "Harmful Tax Competition: An Emerging Global Issue," which focused on very highly mobile flows of financial capital, services, and intangible capital. It specifically recommended against setting very low effective tax rates, but only if the low tax rates were accompanied by at least one other element that would encourage tax competition, such as nontransparency. The report was clearly targeting the establishment of tax havens, among its various goals. More recently, in 2013, the EU proposed moving to a Common Consolidated Corporate Tax Base that would both ease the burden of computing tax liabilities for companies that operate in more than one country and prevent individual countries from enacting special provisions in their corporate tax bases that would promote tax competition. It allows countries to set whatever tax rates they wish, however.

These attempts to limit tax competition are clearly aimed at preventing the kinds of practices that allow MNEs to avoid tax liability rather than preventing direct competition over tax rates. Their effectiveness in limiting tax competition is highly problematic, however, since none of them carry the force of law (Zodrow, 2003; ECTCU).

REFERENCES

European Commission, Taxation and Customs Unions (ECTCU), Common Tax Base, ec.europa.eu/taxation_customs/taxation/company_tax/common_tax_base/index_em.htm.

Gordon, R., Hines Jr., J., 2002. International taxation. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier Science B. V., (Chapter 28).

Hoyt, W., July 1991. Property taxation, nash equilibrium, and market power. Journal of Urban Economics 30 (1), 123–131.

Keen, M., Konrad, K., July 2012. International tax competition and coordination. In: Working Paper 2012-06. Max Planck Institute for Tax Law and Public Finance, Department of Business and Tax Law and Department of Public Economics.

KPMG's Corporate Tax Rate Survey, October 2006. An International Analysis of Corporate Tax Rates from 1993 to 2006, 304-250. KMPG. http://www.kpmg.com/global/en/services/Pages/default.aspx.

Richmond, P., 1963. Taxation of Foreign Investment Income: An Economic Analysis. Johns Hopkins Press, Baltimore.

Wildasin, D., November 1991. Some rudimentary 'Duopoly' theory. Regional Science and Urban Economics 21 (3), 393–421.

Wilson, J., May 1986. A theory of interregional tax competition. Journal of Urban Economics 19 (3), 296–315.

Wilson, J., November 1991. Tax competition with interregional differences in factor endowments. Regional Science and Urban Economics 21 (3), 423–451.

Wilson, J., June 1999. Theories of tax competition. National Tax Journal 52 (2), 269–304.

Wilson, J., Wildasin, D., June 2004. Capital tax competition: bane or boon. Journal of Public Economics 88 (6), 1065–1091.

Zodrow, G., December 2010. Capital mobility and capital tax competition. National Tax Journal 63 (4-Part 2), 865–902.

Zodrow, G., November 2003. Tax competition and tax coordination in the European union. International Tax and Public Finance 10 (6), 651–652.

Zodrow, G., Mieszkowski, P., May 1986. Pigou, Tiebout, property taxation, and the underprovision of local public goods. Journal of Urban Economics 19 (3), 356–370.

Appendix

THE INTENSIVE AND EXTENSIVE MARGINS

The analysis in the text on the labor supply response to taxes and transfers has assumed that people are already working and vary their hours worked as tax and transfer parameters are changed. Economists refer to these responses as occurring along the intensive margin. Another important response, however, is the decision of whether to participate in the labor force at all. A tax will cause some people who are currently working to stop working. Similarly, a wage subsidy will cause some people who are not working to take a job. Economists refer to these participation responses as occurring along the extensive margin. Empirical evidence suggests that responses to taxes and transfers along the extensive margin may be at least twice as strong as responses along the intensive margin for wives, single females, and the young at the low end of the income distribution. Hence the extensive margin has been of particular interest in the design of transfer programs for low-skilled, low-income workers, both the poor and the near poor.

Emmanual Saez has developed a simplified, discrete version of the Mirrlees optimal tax model that is useful for comparing the responses along the two margins. Saez's stripped-down model has the same implications as do more complete models for the design of transfer programs targeted to the low end of the income distribution, and the intuition behind the results is easy to grasp.[1]

THE DISCRETE MODEL

Individual choices and preferences—Individuals choose between I distinct occupations delineated by different, increasing skill levels, $i = 1, \ldots, I$, and unemployment, with unemployment indexed as 0. The different jobs are perfect substitutes in production, such that the wages, w_i, in each job are constant. The wages increase with skill levels. Individuals have preferences defined over the occupation choices, i, and their wage income net of the taxes (+) or transfers (−), T_i, levied by the government on individuals in each occupation i. The net wage income for someone choosing occupation i is $c_i = w_i - T_i$. Therefore, the individuals' utility functions are $U = U(c_i, i)$.

Social welfare—Saez models the individuals as a continuum that divide themselves into $I + 1$ distinct groups based on the occupation they choose, $M_0, M_1, \ldots M_i, \ldots, M_I$. Social welfare is the integral of the individual utilities. Instead of choosing the common Benthamite utilitarian social welfare function, Saez allows for a separate social marginal welfare weight, α^m, for each group. The α^m incorporate society's distributional preferences and give a motivation for transfer payments to people at the low end of the skills/income distribution. Social welfare $W = \int_M \alpha^m U^m(w_i - T_i)\mathrm{d}v(m)$. Lower case m indexes the people in each subgroup M_i.

The government—In addition to redistributing by levying separate taxes or transfers T_i on people within each occupation, the government provides a good allocated to every person. Let H be the per capita amount of the publicly provided good. Saez normalizes the population to one, with h_i defined as the proportion of individuals who choose occupation i on the basis of solving their utility maximization problems. Therefore, $\sum_{i=0}^{I} h_i = 1$. With this normalization, the government's budget constraint is $\sum_{i=0}^{I} h_i T_i = H$. The government's problem, then, is to maximize social welfare with respect to the T_i, subject to the government budget constraint. Forming the Lagrangian

$$\underset{(T_i)}{\text{Max}}\ \mathrm{L} = \int_M \alpha^m U^m(w_i - T_i)\mathrm{d}v(m) + \lambda\left(\sum_{i=0}^{I} h_i T_i - H\right)$$

The FOC with respect to the T_i are

$$-\int_M \alpha^m \frac{\partial U^m(c_i, i^*)}{\partial c_i}\mathrm{d}v(m) + \lambda\left(h_i - \sum_{i=0}^{I} T_i \frac{\partial h_i}{\partial c_i}\right) = 0 \qquad \text{(A.1)}$$

$i = 0, \ldots I, \ldots, I$, with $\delta c_i = -\delta T_i$. The FOC make use of the envelope theorem, which implies that utilities of individuals who are on the margin between groups and who move between groups are unchanged by marginal changes in their wage income net of taxes.

1. Saez also developed a more complete model that incorporates both the intensive and extensive margins in Saez (2000). The model here is from Saez (2002).

Rather than working with the marginal social welfare weights and the utility functions, Saez defines a function $g_i = \frac{1}{\lambda h_i}\int_M \alpha^m \frac{\partial U^m(c_i,i^*)}{\partial c_i}\,\mathrm{d}v(m)$. The Lagrangian multiplier λ is the marginal social welfare of providing one more unit of H to everyone. Since H does not enter into individuals' utility functions, this is equivalent to giving everyone one more dollar of wage income. Therefore, the g_i are ratios of marginal social welfare or social marginal rates of substitution $\left(\frac{\mathrm{d}H}{h_i \mathrm{d}c_i}\right)$, in this case the amount of income given to everyone ($\mathrm{d}H$) that society is willing to give up to give another dollar of income to everyone in occupation i ($h_i \mathrm{d}c_i$). Saez assumes that all the g_i are positive and nonincreasing in i, $i = 0,\ldots, i,\ldots, I$. Lower-income groups have the higher social marginal value, in line with a redistributional motivation.[2]

By the definition of the g_i, the FOC can be written as

$$(1 - g_i)h_i = \sum_{j=0}^{I} T_j \frac{\partial h_j}{\partial c_i} \tag{A.2}$$

Summing over all i

$$\sum_{i=0}^{I}(1 - g_i)h_i = \sum_{j=0}^{I} T_j \sum_{i=0}^{I} \frac{\partial h_j}{\partial c_i}. \tag{A.3}$$

At this point, Saez simplifies the model further by assuming away income effects of the labor supply response, which, at least for transfer programs, appear to be small compared with the substitution effects.[3] Absent income effects, the derivatives $\frac{\partial h_j}{\partial c_i}$ are the Slutsky substitution effects and hence sum to zero. Therefore, $\sum_{i=0}^{I}(1 - g_i)h_i = 0$ and $\sum_{i=0}^{I} h_i g_i = 1$. This result provides a normalization for the g_i and will be used below to characterize the form of the low-income transfer programs under the two margins.

The $i + 1$ FOC and the government's budget constraint solve for the optimal taxes and transfers T_i within each occupation.

THE EXTENSIVE MARGIN

The simplest version of the model to capture the extensive margin is to assume that each individual can work in only one occupation, the one commensurate with his skill level.[4] Therefore, the only choices are to work in that occupation or become unemployed. This implies that the h_i are functions only of c_i and c_0, where c_0 is the subsidy to the unemployed ($w_0 = 0$): $h_i = h_i(c_i, c_0)$.

The FOC for T_i becomes

$$(1 - g_i)h_i = T_i \frac{\partial h_i}{\partial c_i} + T_0 \frac{\partial h_0}{\partial c_i}. \tag{A.4}$$

But $\frac{\partial h_i}{\partial c_i} + \frac{\partial h_0}{\partial c_i} = 0$. Therefore,

$$(1 - g_i)h_i = (T_i - T_0)\frac{\partial h_i}{\partial c_i} \tag{A.5}$$

Equation (A.5) has the following interpretation. The LHS is the effect of increasing taxes by \$1 on everyone remaining in occupation i, which Saez calls the mechanical effect of the tax increase. The dollar raised per person does not have a social marginal value of \$1, however, but only $\$(1 - g_i)$. Because each person in occupation i has one less dollar, the social marginal value declines by g_i dollars by the definition of g_i. Therefore, the net gain in social welfare expressed in terms of dollars is only $\$(1 - g_i)h_i$ for all the people in occupation i.

The RHS is the response effect. With $\frac{\partial h_i}{\partial c_i}$ opting for unemployment, the government collects $(T_i - T_0)$ less in revenue from them, a loss in social welfare in terms of dollars equal to $-(T_i - T_0)\frac{\partial h_i}{\partial c_i}$. The sum of these two effects must be zero at the optimum level of T_i, generating Eqn (A.5).

To express Eqn (A.5) in terms of a response elasticity, define the participation or extensive elasticity as $\eta_i = \frac{\partial h_i}{\partial(c_i - c_0)}\frac{(c_i - c_0)}{h_i}$. Substituting η_i into the RHS of Eqn (A.5), dividing by $(c_i - c_0)$, and rearranging terms, generates

$$\frac{(T_i - T_0)}{(c_i - c_0)} = \frac{1}{\eta_i}(1 - g_i). \tag{A.6}$$

THE INTENSIVE MARGIN

The simplest way to represent the intensive margin in the discrete model is to assume that individuals in occupation i have three choices when the government changes T_i: stay in occupation i or move to one of the two adjacent occupations, $i + 1$ and $i - 1$. They are not constrained to choose only one occupation that matches their skills, as in the case of the extensive margin. These choices imply that $h_i = h_i(c_{i+1}\, c_i, c_{i-1})$.

The FOC for T_i becomes

$$(1 - g_i)h_i = -T_{i+1}\frac{\partial h_{i+1}}{\partial(c_{i+1} - c_i)} - T_i \frac{\partial h_i}{\partial(c_{i+1} - c_i)} + T_i \frac{\partial h_i}{\partial(c_i - c_{i-1})} + T_{i-1}\frac{\partial h_{i-1}}{\partial(c_i - c_{i-1})} \tag{A.7}$$

2. The g_i are endogenous, since they are functions of all the c_i.
3. For an overview of labor supply responses to taxes and transfers along both margins in the empirical literature, see Blundell and MaCurdy (1999). A very good empirical analysis of labor supply responses under the U.S. EITC and public assistance programs is Meyer and Rosenbaum (2001).
4. This is equivalent to assuming that $U^m(c_j, j) = -\infty$, $j \neq i$ because of a skills mismatch.

For any two adjacent movements

$$\frac{\partial h_{i+1}}{\partial(c_{i+1}-c_i)} = -\frac{\partial h_i}{\partial(c_{i+1}-c_i)}$$

Therefore

$$\begin{aligned}(1-g_i)h_i = &-(T_{i+1}-T_i)\frac{\partial h_{i+1}}{\partial(c_{i+1}-c_i)} \\ &+(T_i-T_{i-1})\frac{\partial h_i}{\partial(c_i-c_{i-1})}\end{aligned} \tag{A.8}$$

To simplify further, and bring the discrete model closer to the standard continuous labor supply model with infinitesimal adjustments in hours worked, Saez considers a change in tax dT applied simultaneously to all occupations from i upward to I. This tax changes holds constant the difference in net wage incomes for all occupations except i and $i-1$. Therefore, the only term on the RHS of Eqn (A.8) is $(T_i - T_{i-1})\frac{\partial h_i}{\partial(c_i-c_{i-1})}$. At the same time, however, dT takes tax revenue from people in all occupations from i to I. Therefore, the LHS of Eq. (A.8) becomes $\sum_{j=i}^{I}(1-g_j)h_j$. This is the mechanical effect of the tax revenue gained from the people who remain in their occupations. As above, the social value of the revenue is valued at $(1-g_i)$ for each person in occupation i. The adjustment effect is $(T_i - T_{i-1})\frac{\partial h_i}{\partial(c_i-c_{i-1})}$. The loss in tax revenue is because some people move from occupation i to occupation $i-1$. At the optimum, the sum of the two effects must equal zero

$$\sum_{j=i}^{I}(1-g_j)h_j - (T_i-T_{i-1})\frac{\partial h_i}{\partial(c_i-c_{i-1})} = 0 \tag{A.9}$$

or

$$\sum_{j=i}^{I}(1-g_j)h_j = (T_i-T_{i-1})\frac{\partial h_i}{\partial(c_i-c_{i-1})} \tag{A.10}$$

To express Eqn (A.10) in terms of a response elasticity, define the intensive elasticity as $\xi_i = \frac{\partial h_i}{\partial(c_i-c_{i-1})}\frac{(c_i-c_{i-1})}{h_i}$. Substituting the elasticity into the RHS of Eqn (A.10), dividing by $(c_i - c_{i-1})$, multiplying by h_i, and rearranging terms, yields

$$\frac{\sum_{j=i}^{I}(1-g_j)h_j}{h_i} = \frac{(T_i-T_{i-1})}{(c_i-c_{i-1})}\xi_i \tag{A.11}$$

or

$$\frac{(T_i-T_{i-1})}{(c_i-c_{i-1})} = \frac{\sum_{j=i}^{I}(1-g_j)h_j}{h_i}\frac{1}{\xi_i} \tag{A.12}$$

DESIGNING TAX/TRANSFER PROGRAMS

The optimal tax-transfer policies under the extensive and intensive margins have very different implications for the design of transfer programs targeted to the low end of the income distribution. Begin with the extensive margin.

The extensive margin—Equation (A.6), reproduced here as Eqn (A.13), determines the appropriate transfer program.

$$\frac{T_i - T_0}{c_i - c_0} = \frac{1}{\eta_i}(1-g_i) \tag{A.13}$$

Given that $\sum_{i=0}^{I}h_i g_i = 1$, and that the g_i are positive and nonincreasing, there is some i^* such that $g_i < 1$ for $i > i^*$ and $g_i > 1$ for $i < i^*$. Also, the participation elasticity η_I is positive. Therefore, for people in occupations above i^*, $T_i - T_0 > 0$, and increasing in i. These people should pay taxes that exceed the transfer to the unemployed, T_0. For people in occupations below i^*, $T_i - T_0 < 0$. These people should receive even higher subsidies than the subsidy to the unemployed, T_0 (negative), with the subsidies decreasing as i increases. This implies a two-step program in the style of the U.S. Earned Income Tax Credit:

1. a subsidy to the unemployed (T_0) and,
2. wage subsidies, T_i (negative), greater than the subsidy to the unemployed for those in the lowest skilled occupations that decrease (become less negative) up to the transfer/tax cutoff point i.

The higher skilled people, those in occupations $i > i^*$, pay the entire cost of these subsidies.

Offering higher subsidies to low skilled people has two advantages. It encourages some of the unemployed to work, which raises the supply of labor, and it gives additional subsidies to people with low skills ($i < i^*$) that have higher social marginal value ($g_i > 1$) than the taxes required to pay for them.

The intensive margin—Equation (A.12), Reproduced here as Eqn (A.14), determines the appropriate transfer program

$$\frac{(T_i-T_{i-1})}{(c_i-c_{i-1})} = \frac{\sum_{i}^{I}(1-g_i)h_i}{h_i}\frac{1}{\xi_i} \tag{A.14}$$

Recall that, absent income effects, $\sum_{i=0}^{I}(1-g_i)h_i = 0$ and the g_i are nonincreasing. Thus for any $i > 0$, $\sum_{j=i}^{I}(1-g_j)h_j > 0$. Since the intensive elasticity ξ_i is positive, $T_i - T_{i-1} > 0$. All marginal tax rates must be positive. This implies a transfer program in the style of a negative income tax (NIT), in which the unemployed receive a guaranteed income and then that income is taxed away beginning with people in the lowest skilled occupations. The marginal tax rates rise with i since

$\sum_{j=i}^{I}(1-g_j)h_j$ becomes more positive the higher is i up to i^*.

For additional insight into the nature of the guaranteed income, consider the optimal taxes/transfers at the bottom of the distribution, T_1 and T_0

$$\sum_{i=1}^{I}(1-g_i)h_i = \sum_{i=1}^{I}h_i - \sum_{i=1}^{I}g_ih_i = (1-h_0) - (1-g_0h_0) = h_0(g_0-1). \tag{A.15}$$

Therefore

$$\frac{(T_1-T_0)}{(c_1-c_0)} = \frac{(g_0-1)h_0}{h_1}\frac{1}{\xi_i} \tag{A.16}$$

The larger the social marginal value of the unemployed, g_0, the higher the tax rate on the people in the lowest skilled occupation, which implies a high phase-out rate of the guaranteed at the bottom of the distribution. This is essentially what U.S. programs such as SNAP (Food Stamps), TANF (Temporary Assistance to Needy Families), and SSI (Supplemental Security Income) do. Because only those who are unemployed receive pure subsidies, an NIT can offer a higher guaranteed income than can the EITC-style program implied by the extensive margin, since the latter has to finance subsidies to a whole range of low-skilled workers.

In contrast, offering EITC-style wage subsidies under the intensive margin has ambiguous implications. It has the benefit of inducing some of the unemployed to work in occupation 1, which raises the supply of labor. But it also induces some people working in occupation 2 to move into occupation 1, which in effect reduces the supply of labor. It does not have the unambiguous benefits that appear under the extensive margin.

Combining the extensive and intensive margins—Saez combines the extensive and intensive models above to generate the following optimal tax formula for T_i that incorporates the two elasticities:

$$\frac{(T_i-T_{i-1})}{(c_i-c_{i-1})} = \frac{\sum_{j=i}^{I}\left(1-g_j-\eta_j\frac{T_j-T_0}{c_j-c_0}\right)h_j}{h_i}\frac{1}{\xi_i} \tag{A.17}$$

Notice that it has the same form as the optimal tax formula for the intensive model with

$g_j - \eta_j\frac{T_j-T_0}{c_j-c_0}$ replacing g_j. In the combined model:

the mechanical effect is: $\sum_{j=i}^{J}(1-g_j)h_j$

the intensive response is: $-\frac{(T_i-T_{i-1})}{(c_i-c_{i-1})}h_i\xi_i$

and the extensive response is: $-\sum_{j=i}^{I}\frac{T_j-T_0}{c_j-c_0}h_j\eta_j$

The three sum to zero at the optimum, generating Eqn (A.17).

Saez runs simulations for the U.S. economy using the combined model. He believes the most representative elasticities are $\eta = 0.5$ and $\xi = 0.25$. He also chooses a pattern for the g_i that has quite a strong distributional motivation: as income increases by a factor of N, the marginal social value of income declines by the same factor N. This may well be more redistributive than U.S. citizens would advocate, as discussed in Chapter 4 of this text. The earnings distribution is based on the March 1997 Current Population Survey. The simulation results imply having a modest guaranteed income, combined with a zero marginal tax rate at the bottom of the skills distribution and then substantial marginal tax rates taxing away the guaranteed income further up the skills distribution.[5] The modest guaranteed income and the zero marginal tax rate at the bottom of the distribution are suggestive of an EITC-type program, whereas the positive marginal tax rates beyond the lowest income category are more in keeping with an NIT-type program.

A TRADE-OFF BETWEEN EFFICIENCY AND EQUITY?

Paying attention to the extensive margin has paid dividends in understanding labor supply responses to both transfers and taxes. As noted earlier, empirical analysis of the labor supply responses to transfers to the poor have suggested that the majority of the response is on the extensive margin.[6] The labor supply responses along the extensive margin to changes in tax rates may have even more dramatic implications.

The mainstream perspective on the maximization of social welfare through taxes and transfers is that it is a negative sum game—Okun's Leaky Bucket described in Chapter 4. Attempts to redistribute to achieve more equality move society underneath its production possibilities frontier because of the inefficiencies associated with taxes and transfers. But an analysis by Rolf Aaberge, Ugo Columbino, and Steiner Strom (ACS) of the labor supply responses of married couples in Italy, Norway, and Sweden to changes in wages suggests that equity and efficiency may not necessarily bear a trade-off relationship after all.

We will consider their results for the Italian couples by way of illustration. They estimated the responses along both the intensive and extensive margins of both the men and women within the poorest 10% of the households, the richest 10% of the households, and the 80% in the middle

5. We leave the details of this section to the Saez article for the interested reader. The simulation results are reported on p. 1063.
6. See, for example, the analysis of the labor supply responses to the EITC by Meyer and Rosenbaum (2001).

of the distribution. The elasticities were near zero for both the richest and poorest men along both margins. The women behaved differently, however. The elasticities remained quite low for the women along the intensive margin, 0.01 for the richest women and 0.47 for the poorest women. The hours worked by the poorest women are more responsive than for the richest women, but their response is still quite inelastic. The difference along the extensive margin was extremely large, however: an elasticity of 0.03 for the richest women, compared with an elasticity of 2.83 for the poorest women, a highly elastic response. ACS conjecture that since the rich tend to marry the rich and the poor tend to marry the poor, poor married women are more likely than rich married women to join the labor force if wages rise simply because their households have a greater need for income.

The implication of their results is that the Italian government, by cutting marginal income tax rates on the poor and raising the rates on the rich, could achieve more equality and reduce the inefficiency of their income tax while raising the same amount of revenue. The efficiency gains arise by switching the tax burden from the highly elastic poor to the highly inelastic rich. Equality and efficiency bear a direct relationship to one another, primarily because of the very different labor supply responses of rich and poor married women along the extensive margin.[7] Since the intensive and extensive elasticities were also quite low for the middle 80% of the couples, an EITC-style program would work well, in line with the Saez analysis above: Provide subsidies to the poorest couples to encourage the women to supply more labor, and then have fairly low tax rates in the phase out region so that the highest marginal tax rates are levied on higher income couples to maximize the redistributional impact of the income tax. The gains would presumably be even greater if the subsidies at low incomes and the higher marginal tax rates at high incomes could be targeted just to the women, but this is almost certain to offend people's sense of horizontal equity. It will be interesting to see if such different labor supply responses between rich and poor women (or men) along the extensive margin are found in other countries as well.

The one caveat according to ACS is that studies have found that high-income workers tend to choose higher paying jobs in response to reductions in their marginal tax rates. If wages and productivity are positively correlated, as they should be, then raising taxes on the rich may reduce the overall productivity of the economy. But the idea that equity and efficiency may not necessarily bear a trade-off relationship in reforming an income tax is intriguing nonetheless.

REFERENCES

Aaberge, R., Colombino, U., Strom, S., December 2000. Labor supply responses and welfare effects from replacing current tax rules by a flat tax: empirical evidence from Italy, Norway, and Sweden. Journal of Population Economics 13 (4), 595–622.

Blundell, R., MaCurdy, T., 1999. Labor supply: a review of alternative approaches. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. IIIA. North-Holland, Amsterdam.

Meyer, B., Rosenbaum, D., August 2001. Welfare, the earned income tax credit, and the labor supply of single mothers. The Quarterly Journal of Economics 116 (3), 1063–1114.

Saez, E., 2000. Optimal Income Transfer Programs: Intensive versus Extensive Labor Supply Responses. NBER Working Paper. No. 7708.

Saez, E., August 2002. Optimal income transfer programs: intensive versus extensive labor supply responses. The Quarterly Journal of Economics 117 (3), 1039–1073.

7. Aaberge et al. (2000). The elasticities are reported in Table 4, p. 604. The pattern of intensive and extensive elasticities was the same for Norway, and the intensive elasticities the same for Sweden. ACS did not estimate the extensive elasticities for the Swedish couples because such a large percentage of married women are already working in Sweden. Note that the estimated elasticities are actual rather than compensated elasticities, but if the income effects on labor supply are fairly small then the large differences in the female responses on the extensive margin are likely to reflect large differences in the compensated elasticities.

Index

Note: Page numbers followed by "f" and "t" denote figures and tables, respectively

C

D

E

F

G

H

P

Q

R

S

T

U

V

W

Y

Z