HANDBOOK OF PUBLIC ECONOMICS

Volume 3

Alan J. Auerbach &
Martin Feldstein

EDITORS' INTRODUCTION TO VOLUME 3

The publication of Volume 3 and the forthcoming Volume 4 of the Handbook of Public Economics affords us several opportunities: to address lacunae in the original two volumes of this series, to revisit topics on which there has been substantial new research, and to address topics that have grown in importance. Indeed, many of the papers individually encompass all three of these elements. For each chapter related to one from an earlier volume, the new contribution is free-standing, written with the knowledge that the reader retains the opportunity to review the earlier chapter to compare perspectives and consider material that the current author has chosen not to cover. Indeed, such comparisons illuminate the evolution of the field during the roughly two decades that have elapsed since work first began on the chapters in Volume 1. Taken together, the four volumes offer a comprehensive review of research in public economics, in its current state and over the past few decades, written by many of the field's leading researchers.

EDITORS' INTRODUCTION

The field of Public Economics has been changing rapidly in recent years, and the sixteen chapters contained in this Handbook survey many of the new developments. As a field, Public Economics is defined by its objectives rather than its techniques, and much of what is new is the application of modern methods of economic theory and econometrics to problems that have been addressed by economists for over two hundred years. More generally, the discussion of public finance issues also involves elements of political science, finance and philosophy. These connections are evidence in several of the chapters that follow.

Public Economics is the positive and normative study of government's effect on the economy. We attempt to explain why government behaves as it does, how its behavior influences the behavior of private firms and households, and what the welfare effects of such changes in behavior are. Following Musgrave (1959) one may imagine three purposes for government intervention in the economy: *allocation,* when market failure causes the private outcome to be Pareto inefficient, *distribution,* when the private market outcome leaves some individuals with unacceptably low shares in the fruits of the economy, and *stabilization,* when the private market outcome leaves some of the economy's resources underutilized. The recent trend in economic research has tended to emphasize the character of stabilization problems as problems of allocation in the labor market. The effects that government intervention can have on the allocation and distribution of an economy's resources are described in terms of efficiency and incidence effects. These are the primary measures used to evaluate the welfare effects of government policy.

The first chapter in this volume, by Richard Musgrave, presents an historical development of these and other concepts in Public Finance, dating from Adam Smith's discussion in *The Wealth of Nations* of the role of government and the principles by which taxes should be set. The remaining chapters in the Handbook examine different areas of current research in Public Economics.

Analyses of the efficiency and incidence of taxation, developed in Musgrave's chapter, are treated separately in Alan Auerbach's chapter in the first volume and Laurence Kotlikoff's and Lawrence Summers' chapter in the second volume, respectively. Auerbach surveys the literature on excess burden and optimal taxation, while Kotlikoff and Summers discuss various theoretical and empirical approaches that have been used to measure the distributional effects of government tax and expenditure policies.

These general analyses of the effects of taxation form a basis for the consideration of tax policies in particular markets or environments, as is contained in the chapters by Jerry Hausman, Agnar Sandmo, Avinash Dixit, Harvey Rosen, John Helliwell and Terry Heaps, and Joseph Stiglitz.

Hausman discusses the effects of taxation on labor supply, including a treatment of how one empirically estimates such effects in the presence of tax and transfer programs. He also considers the incentive effects of social welfare programs such as unemployment compensation and social security. Sandmo focuses on the other major factor in production, capital, dealing with theory and evidence about the effects of taxation on private and social saving and risk-taking. Dixit shows how the basic results about the effects of taxation may be extended to the trade sector of the economy, casting results from the parallel trade literature in terms more familiar to students of Public Finance. Rosen's chapter brings out the characteristics of housing that make it worthy of special consideration. He considers the special econometric problems involved in estimating the response of housing demand and supply to government incentives. Because of its importance in most family budgets and its relatively low income elasticity of demand, housing has been seen as a suitable vehicle for government programs to help the poor, and Rosen discusses the efficiency and incidence effects of such programs. Helliwell and Heaps consider the effects of taxation on output paths and factor mixes in a number of natural resource industries. By comparing their results for different industries, they expose the effects that technological differences have on the impact of government policies. Stiglitz treats the literature on income and wealth taxation.

The remaining chapters in the Handbook may be classified as being on the "expenditure" side rather than the "tax" side of Public Finance, though this distinction is probably too sharp to be accurate. In Volume 1, Dieter Bös surveys the literature on public sector pricing, which is closely related both to the optimal taxation discussion in Auerbach's chapter and Robert Inman's consideration, in Volume 2, of models of voting and government behavior. The question of voting and, more generally, public choice mechanisms, is treated by Jean-Jacques Laffont in his chapter.

The chapters by William Oakland and Daniel Rubinfeld focus on the provision of "public" goods, i.e., goods with sufficiently increasing returns to scale or lack of excludability that government provision is the normal mode. Oakland considers the optimality conditions for the provision of goods that fall between Samuelson's (1954) "pure" public goods and the private goods provide efficiently by private markets. Rubinfeld surveys the literature on a special class of such goods: local public goods. Since the work of Tiebout (1956), much research has been devoted to the question of whether localities can provide efficient levels of public goods.

The other two chapters in Volume 2 also deal with problems of public expenditures. Anthony Atkinson considers the effects of the range of social welfare programs common in Western societies aimed at improving the economic standing of the poor. Some of these policies are touched on in the chapters by Hausman and Rosen, but the coexistence of many different programs itself leads to effects that cannot be recognized

by examining such programs seriatim. Jean Drèze and Nicholas Stern present a unified treatment of the techniques of cost benefit analysis, with applications to the problems of developing countries.

References

Musgrave, R.A. (1959), The Theory of Public Finance (McGraw-Hill, New York).
Samuelson, P.A. (1954), "The pure theory of public expenditures", Review of Economics and Statistics 36:387–389.
Tiebout, C.M. (1956), "A pure theory of local expenditures", Journal of Political Economy 94:416–424.

INTRODUCTION TO THE SERIES

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

KENNETH J. ARROW and MICHAEL D. INTRILIGATOR

PUBLISHER'S NOTE

For a complete overview of the Handbooks in Economics Series, please refer to the listing at the end of this volume.

CONTENTS OF THE HANDBOOK

VOLUME 1

VOLUME 2

THE TRANSFORMATION OF PUBLIC ECONOMICS RESEARCH: 1970–2000

MARTIN FELDSTEIN*

The nature and content of research and teaching in public economics have changed enormously during the past three decades. The field is more theoretically rigorous, more empirical, more focused on real policy issues, and more concerned with government spending as well as with taxation. For me, it has been an exciting time to be a public finance economist and to contribute to this intellectual transformation.

Theoretical beginnings

When I began studying public finance as a graduate student in England in the early 1960s, the bible of the field was Richard Musgrave's *The Theory of Public Finance* (1959). Unlike earlier books by authors like Pigou (1947) which were characterized by prose unencumbered by diagrams and algebra, most of the Musgrave volume looked like a standard price theory book with graphs and algebra showing the partial equilibrium effects of taxes on prices and quantities and the associated effects on deadweight losses. The Musgrave book was about the core issues of incidence and efficiency and the positive effects on the actions of buyers and sellers without the detailed descriptions of tax rules or administrative issues that characterized many earlier public finance books. Although this text opened up a new era in public finance, its limited mathematics meant that it was weak in dealing with multi-product problems and in analyzing general equilibrium effects. The general absence of references to econometric research reflected the state of the field at the time. Similarly, although Musgrave discussed general principles of government spending, his classic text did not deal with the specific areas of government spending that would become the subject of much of public economics in the past three decades.

Arnold Harberger's work on the incidence of the corporate income tax [Harberger (1962)] demonstrated the power and importance of simple general equilibrium

* Professor of Economics, Harvard University and President of the National Bureau of Economic Research. This essay, written for the 30th Anniversary of the *Journal of Public Economics,* focuses on research that has been described in the English language and therefore primarily on work done in the United States and Britain. The *Journal* played an important role in the transformation described here. I am grateful for the opportunity to serve as a Co-Editor of the *Journal* from 1972 through 1986, as an associate editor from 1987 through 1997, and as an advisory Editor since that time.

models. By extending models originally developed to study international trade issues, Harberger showed how elasticities of substitution in production and consumption, factor intensities in production, and consumer preferences all combined to determine the incidence of the corporate tax on labor and capital and on consumers with different preferences. Gone were the earlier vague statements about backward shifting and forward shifting. Although the new general equilibrium models did not give unambiguous answers about corporate tax incidence, we learned the reason for the ambiguity and how various factors like capital mobility would affect incidence.

In two further studies Harberger (1964, 1966) showed how the traditional welfare loss triangle could be extended to multiple taxes on different products and to evaluating the deadweight loss of the corporate income tax. Although multi-product deadweight loss calculations had been developed earlier by Irving Fisher (1937), and John Hicks (1939), it was Harberger who showed their direct application to excise taxes. Corlett and Hague (1953) made a seminal contribution to the theory of the efficient design of multi-product excise taxes when some products are non-taxable or are taxed at an arbitrary rate. With these ideas well established, the growing mathematical literacy of the economics profession led to a rediscovery of the Frank Ramsey's (1927) theory of optimal excise taxes. Diamond and Mirrlees (1971) modernized Ramsey's analysis, showed the optimality of maintaining production efficiency, and derived the conditions that generalized the traditional inverse elasticity rules for optimal taxation.

At about the same time, Mirrlees (1971) also developed a formal model of the optimal labor income tax in which the optimal degree of progressivity depends on the government's distributional preferences and on the responsiveness of individuals to the tax schedule. The research provided a formal structure for guiding a benevolent government through the process in which the government optimizes the schedule of income tax rates knowing that the taxpayers will respond by maximizing their own utility subject to the schedule of tax rates. Although the analysis failed to provide any significant general results, it clarified the nature of the optimization problem and provided a framework for deriving results in models with more explicit parametric restrictions.

A further generalization of the original Diamond–Mirrlees analysis dealt with designing the optimal combination of income and excise taxes. In the end, that research showed that the optimal tax rules depend on such unobservable properties of the utility function as the separability between leisure and the components of consumption as well as on the higher derivatives of utility as a function of income.

These theoretical developments led to other studies of tax incidence in general equilibrium models (including the important early work on computable general equilibrium analysis by John Shoven and John Whalley), to extensions of the Diamond–Mirrlees optimal tax analysis to include expenditure issues, to new work on the incidence of taxes on corporate source income by David Bradford, Mervyn King, and others, and to my own research on the efficiency effect of taxes on capital income.

These developments in the theory of public finance in the 1960s and 1970s were important in two ways. First, they clarified enormously the profession's thinking about a number of important public finance questions. Although they did not give unambiguous

answers, they showed the errors of some earlier views and provided substantial analytic insights. Second, they attracted an outstanding generation of graduate students to the field of public economics. Most of them did not go on to do theoretical research but the improved theoretical foundations in public finance and the new standard of theoretical rigor contributed to their empirical work.

Empirical research

The development of empirical work in public economics has, more than anything else, distinguished the research of the past 30 years from all that had gone before. The late 1960s and early 1970s saw for the first time the availability of high-speed computers, reliable econometric software, and large machine-readable data sets. These developments, plus the addition of sophisticated econometric techniques to the standard tool kit of graduate students, were all key to the empirical revolution in public economics.

The new data for public finance research included the first public availability of the Current Population Survey, the Federal Reserve's Survey of Consumer Finances, and the Internal Revenue Service public use sample of 100 000 tax returns that became the basic data input for what is now the NBER Taxism model. For someone like me, recently trained in econometric methods, the newly available data provided an exciting opportunity to do a kind of empirical public finance that had not been done before and to confront some of the key questions of public finance in a new and serious empirical way.

An important early subject of empirical research was the study of the effects of taxes on labor supply, or, more accurately, on labor force participation and hours. These studies benefited also from new econometric techniques for dealing with limited dependent variables and with self-selection bias in estimating behavior from a subset of the population. The results showed important effects of taxes on the participation and hours of women. But the apparent lack of response by men was a warning that an accurate characterization of labor supply must be a much broader measure that includes things like effort, location, acquisition of human capital, and choice of occupation.

More generally, what matters for evaluating the deadweight loss of the distortions induced by labor income taxes is not the change in labor supply alone (even broadly defined) but the change in the individual's taxable income, including the effect on the form of compensation (i.e., on the choice of fringe benefits and working conditions instead of cash) and on the deductions taken by individuals who itemize their tax returns [Feldstein (1999)]. Fortunately, unlike the impossibility of studying broadly defined labor supply, it is possible to estimate the effect of changes in marginal tax rates on taxable income using panels of tax data that include repeated observations on the same individuals or, under some conditions, using pooled cross sections of data.

Econometric tax research on the effect of interest income taxes on household saving is difficult because neither the tax return panels nor other panel files give adequate data

on saving. Time series data on saving indicates that taxes that reduce the net return on saving do depress saving but these results are subject to a variety of estimation problems. Much more solid evidence on the effects of tax policies on saving have been derived in studies of the effects of IRA and 401-k plans. Although controversy continues, the evidence appears to support the conclusion that these saving incentives do significantly increase overall saving.

A series of legislative changes in the tax treatment of capital gains provided the basis for several studies of the effect of the capital gains tax rate on the selling of corporate stock and the realization of capital gains. Related studies analyzed how tax rates affect the way households allocate their wealth among different types of financial assets. Although results differ among the individual studies, the overall implication is that households do respond to differences in tax rates and to changes in tax rules.

Closely related to these studies of the effect of taxes on financial investment are the studies of the effects of marginal tax rates on home ownership. Because mortgage interest payments are deductible in calculating taxable income while the imputed value of housing services is not included in taxable income, individuals with high marginal tax rates have a strong incentive to own a home and to increase their investment in owner-occupied housing when tax rates rise. Several econometric studies confirm that both inferences are correct, estimate the magnitude of the distortion, and calculate the resulting efficiency losses. Other empirical studies of the effects of taxation deal with such things as charitable giving and the demand for health insurance. There is, in short, no aspect of household tax-related behavior that has not been studied. But with new tax policies and improved data sets, there will be new opportunities in the future to improve and refine our empirical knowledge in a wide range of areas.

In addition to these empirical studies, there have also been analytic studies of taxation that sharpened our understanding of the effect of taxes on risk taking by individuals, of how taxes affect the financial policy of corporations, and of the implications of analyzing tax issues in the context of a growing economy.

Government spending

A second major aspect of the transformation of research in public finance since the 1960s has been to broaden the subject to include government spending as well as taxation. This shift in focus was no doubt stimulated by the enormous expansion of government spending. In the United States, non-defense spending of the federal government rose from less than 10% of GDP in 1965 to more than 15% in 2000, reflecting a wide array of new programs ranging from pre-school education to health care for the aged. Economists responded to the challenge of studying these new programs. The field of public finance was thus transformed from the study of the taxes used to finance basic government services to the field of public economics that looked also at the effect of government spending on a wide range of programs.

Much of the growth of government spending has been for social insurance programs and the research in public economics has matched that emphasis. Social Security pensions, unemployment insurance, workers' compensation, and the Medicare/Medicaid programs of health care for the aged and the poor raised new theoretical as well as empirical issues that became a major focus for research.

Social insurance programs were attractive research subjects not only because they are the largest part of government spending but also because they have many analytic similarities to taxation. The analyses of public spending programs study not only the extent to which they achieve their stated purposes but also the incidence and excess burden of each program. The design of social insurance programs involves tradeoffs between protection and distortion that are analogous to the tradeoffs between distribution and efficiency considerations in taxation.

The Social Security program of benefits to retirees, dependants and the disabled is the largest form of government spending. Empirical studies have shown that Social Security reduces saving and induces early retirement in the United States and other countries. In addition to these studies of the behavioral effects of Social Security, there have been a variety of empirical studies of the general equilibrium effects of Social Security and Social Security reform, including the effects of shifting from the current pay-as-you-go system to systems that rely in whole or in part on investment-based accounts. Separately, analytic studies have examined the optimal design of social security retirement and disability programs.

Studies of other social insurance programs, including disability insurance and workers' compensation, also estimated behavioral effects, analyzing the distortions to incentives and the efficacy of the programs in providing the protection for which they are intended.

Government health care programs are important fiscally as well as socially. Even in the United States, the government accounts for nearly half of total health care spending and exceeds six percent of GDP. The large volume of microeconomic data about the cost and provision of health care services also encouraged the growth of research in this area. The introduction of changes in the state level Medicaid program at different times in different states provided a source of identification for studying different aspects of this significant program.

While early work on the economics of education focused on measuring and explaining the returns to human capital accumulation, the public finance research on education has looked at issues like the effect of alternative local tax and grant rules on the level and distribution of local government education spending. Important also have been the Tiebout-inspired analyses of the effects of competition in education on various educational outcomes. The government's increased role in providing scholarships for higher education has also induced public finance economists to study the impact of such spending on college enrolment and graduation as well as on household saving.

Other government programs ranging from child care to the criminal justice system have been the subject of public finance studies that compare the cost of achieving

program goals to the full cost of the taxes needed to finance that spending, including the deadweight loss associated with that tax revenue.

Macroeconomic issues

Although Keynesian fiscal policy was a major focus of Richard Musgrave's *The Theory of Public Finance,* by the 1970s the analysis of stabilization policies had largely shifted to the field of macroeconomics where the emphasis was much more on monetary policy than on variations in fiscal stimulus through changes in budget deficits and surpluses. Public finance research nevertheless contributed to the debate by studying how tax rules like the investment tax credit and depreciation allowances could be used to stimulate business investment in a counter-cyclical way.

The major social insurance programs also lie on the border between macroeconomics and public finance. Unemployment insurance raises the level of unemployment and contributes to its cyclical volatility. Public finance researchers have crossed the border into macroeconomics and labor economics to study the effect of unemployment insurance on the level and character of unemployment and to analyze ways in which unemployment insurance can be improved to reduce the inefficient labor market distortions without decreasing protection against the hardships of unemployment. Social Security pensions can also have an important macroeconomic effect by changing the rate of capital accumulation and therefore the rate of economic growth.

The high inflation rate in the late 1970s inspired research on how the interaction of inflation and tax rules affects the level and distribution of saving and investment. This research showed that the neutrality of money and money growth in theoretical macroeconomic models does not hold in actual economies that tax nominal capital income. The analysis also led to calculations showing that the substantial deadweight loss of even moderate rates of inflation.

Fiscal federalism

The complex federal structure of the U.S. government assigns important decision-making authority to state and local governments. Those state and local governments are now responsible for spending an amount equal to more than 60% of the spending by the federal government. An important area of public economics research has been the analysis of how those governments choose their tax and spending policies, how those choices are affected by the policies of higher levels of government (including block grants and matching grants), and how the resulting inter-area differences in taxes and spending affect the behavior of the private sector and the outcomes of government programs. Although such work has dealt primarily with the United States, it is likely to become more important as the European Union evolves toward a more federal fiscal structure.

Looking ahead

The past three decades have been an enormously productive period for the field of public economics with important advances in theoretical analysis and empirical knowledge. The central role of the government in the economy and the associated high marginal tax rates mean that the problems of taxing and spending will continue to provide challenging opportunities for research in public economics. If those studies are to be useful in improving public policy, they must continue to speak to the real problems of the economy and must combine appropriate analytic models with sound empirical research.

Cambridge, MA
September 2001

References

Bradford, D. (1981), "The incidence and allocation effects of a tax on corporation distributions", Journal of Public Economics 15:1–22.

Corlett, W.J., and D.C. Hague (1953), "Complementarity and the excess burden of taxation", Review of Economic Studies 21:21–30.

Diamond, P.A., and J.A. Mirrlees (1971), "Optimal taxation and public production I: production efficiency; II: tax rules", American Economic Review 61:8–27; 261–278.

Feldstein, M. (1999), "Tax avoidance and the deadweight loss of the income tax", Review of Economics and Statistics 81(4):674–680.

Fisher, I. (1937), "Income in theory and income taxation practice", Econometrica 5:1–55.

Harberger, A.C. (1962), "The incidence of the corporation income tax", Journal of Political Economy 70:215–250.

Harberger, A.C. (1964), "Taxation, resource allocation, and welfare", in: J. Due, ed., The Role of Direct and Indirect Taxes in the Federal Revenue System (Princeton University Press, Princeton, New Jersey).

Harberger, A.C. (1966), "Efficiency effects of taxes on income from capital", in: M. Krzyaniak, ed., Effects of Corporate Income Taxes (Wayne State University Press, Detroit).

Hicks, J.R. (1939), Value and Capital (Oxford University Press, London).

King, M.A. (1977), Public Policy and the Corporation (Chapman and Hall, London).

Mirrlees, J.A. (1971), "An exploration in the theory of optimum income taxation", Review of Economic Studies 38:175–208.

Musgrave, R.A. (1959), The Theory of Public Finance (McGraw-Hill, New York).

Pigou, A.C. (1947), A Study in Public Finance, 3rd edition (Macmillan, London).

Ramsey, F.P. (1927), "A contribution to the theory of taxation", Economic Journal 37:47–61.

Shoven, J.B., and J. Whalley (1972), "A general equilibrium calculation of the effects of differential taxation income from capital in the U.S.", Journal of Public Economics 1:281–321.

Chapter 17

TAXATION, RISK-TAKING, AND HOUSEHOLD PORTFOLIO BEHAVIOR*

JAMES M. POTERBA
MIT and NBER

Contents

* I am grateful to Alan Auerbach, Scott Weisbenner, and participants at the Burch Center Symposium for helpful comments, to Andrew Mitrusi for providing me with data from TAXSIM, and to the Hoover Institution and National Science Foundation for research support.

Handbook of Public Economics, Volume 3, Edited by A.J. Auerbach and M. Feldstein

Abstract

This chapter summarizes the current state of research on how taxation affects household decisions with respect to portfolio structure and asset trading. It discusses long-standing issues, such as the impact of differential taxation of income flows from stocks and bonds on the incentives for households to invest in these assets, and the effect of capital gains taxation on asset sales. It also addresses a range of emerging issues, such as the impact of taxation on the behavior of mutual funds and their investors, and the effect of tax changes and tax uncertainty on investor behavior. It concludes that taxation exerts a systematic influence on the nature of risk-taking and the structure of household portfolios. Research on the effects of taxation on portfolio structure is more advanced than work on the welfare costs of portfolio distortions.

Keywords

portfolio choice, after-tax returns, investor behavior

JEL classification: H42, G11

Introduction

How taxation affects household saving is one of the most-studied issues in empirical public finance. The reason for interest in this issue is clear: the supply of saving is a key determinant of the cost of capital and therefore of the amount of productive investment in an economy. By comparison, the effect of taxation on the allocation of household saving across different asset categories has received far less research attention. This is surprising, since the supply of funds to particular sectors can be just as important as the overall level of saving in determining the cost of capital for particular types of investment.

This survey considers the existing state of research on how taxation affects risk-taking, portfolio choice, and the allocation of household saving. It describes both the theoretical models that have describe how optimizing households might allocate their portfolio holdings across different assets, as well as empirical evidence that explores the link between taxation and portfolio structure. The chapter considers both decisions about which assets to hold, as well as decisions about when to sell particular assets. The chapter also discusses a number of emerging issues concerned with taxation and portfolio structure, such as the effect of taxation on mutual fund investors and investors who take advantage of tax-deferred investment vehicles such as Individual Retirement Accounts and 401(k) plans in the United States.

The chapter is divided into six sections. The first presents a brief overview of the taxation of capital income in developed countries, with particular focus on the current tax rules in the United States. This includes a discussion of the aggregate household balance sheet, as a way of introducing the relative importance of different assets that households own. Section 2 considers the effect of taxation on the set of assets held in household portfolios, and the portfolio shares held in different assets. It begins with the traditional theory of taxation and the demand for risky assets. It then considers the impact of differential taxation of different types of capital income on the demand for assets that provide different income flows, for example, on corporate stocks that provide returns in the form of dividends rather than capital gains.

Section 3 explores the effect of taxation on capital asset sales, with particular attention to the link between capital gains tax rates and the decision to realize gains. It also considers the potential effect of securities-transaction taxes on financial markets and investor behavior. The fourth section explores a variety of topics related to taxation and portfolio choice, including taxation and investment in mutual funds, taxation and life-insurance products, and the role of estate taxation in affecting portfolio choice. Section 5 considers the link between taxation and investment decisions in human capital. The sixth section concludes and raises a number of issues concerning taxation and portfolio choice that require further investigation.

1. Taxation and the portfolio choice environment

How the tax system affects risk-taking and portfolio choice depends on a number of different provisions in the tax code and on the set of financial assets that are available

to investors. The tax rules that apply to income from capital are the most complicated part of most modern income tax systems. The income from different types of capital assets may be taxed at different rates, different types of income from the same asset may be taxed at different rates, and different investors may face different tax rates on the same asset. In addition, there are substantial differences across countries in both the level and structure of capital income taxes.

Many of the tax provisions that affect the after-tax returns from different assets are straightforward to summarize, but it is more difficult to describe their ultimate impact on the high-net-worth households who account for a substantial fraction of aggregate net worth. Poterba (2000a) reports information from the 1998 Survey of Consumer Finances on the concentration of net worth. The households in the top 0.5 percent of the net worth distribution hold 26 percent of net worth. For some asset categories, such as publicly traded corporate stock excluding ownership through pensions or mutual funds, the concentration of ownership is even greater. Forty-one percent of directly-held stock is held by the households in the top 0.5 percent of the ownership distribution; over 80 percent is held by households in the top five percent.

Analyzing the effect of taxation on high-net-worth households is difficult because these households typically receive sophisticated tax advice, and they may find strategies that enable them to avoid the tax burdens associated with simple application of the tax statutes. The fees of their tax planning advisers, and the pre-tax returns that they forego to maximize after-tax returns, represent implicit taxes on their capital income. The need to recognize such implicit taxes and to consider their distortionary effects is one of the central themes of Scholes, Wolfson, Erickson, Maydew and Shevlin (2002).

While it is straightforward to describe the set of financial assets that are potentially available to investors, it is often difficult to calculate the transaction costs that are associated with holdings of these assets, and therefore the set of assets that are available at reasonable cost to many investors. Moreover, because wealth data are one of the most difficult types of survey data to collect, for many nations there is relatively little information on the composition of household-net-worth and the structure of household portfolios.

This section illustrates the taxation of portfolio income and the analysis of household portfolio structure by focusing on the United States. It begins with a discussion of the tax rules on investment income and then considers the current structure of household portfolios.

1.1. The taxation of investment income

Most developed nations tax interest income, and many also tax dividends received by individuals. OECD (1994) provides a valuable introduction to the tax rules on capital income in a range of developed nations.

In the United States, individuals are taxed at equal rates on their dividend and interest income, and the personal-income tax on dividend income is not integrated with the corporation tax. In addition to federal income tax, state and (in some cases) local

income taxes may also apply to interest and dividend receipts. For calendar year 2000, the maximum statutory federal income tax rate was 39.6 percent, although the effects of various exemption phase-outs could increase the marginal tax rate to between 40 and 42 percent. Mitrusi and Poterba (2001) show that very few taxpayers in the "top bracket" actually face the 39.6 percent rate; far more face rates over 40 percent. State income tax rates can substantially increase the total tax burden on investment income. In New York, one of the highest tax-rate states, the top personal income tax rate exceeds 10 percent. The effective state income tax rate is reduced somewhat because these taxes can be deducted from federal taxable income, but even with this deduction, the top marginal tax rate on dividend and capital gains income is currently near 50 percent. Shackelford (2000a) discusses the tax rules facing high-net-worth households in more detail.

Realized capital gains are also taxed in the United States, although they are not taxed in all developed nations. Most conceptual discussions of comprehensive income taxation focus on accrued rather than realized gains as a part of the tax base; the practical difficulties of taxing accrued gains has led essentially all nations that tax capital gains to tax them at realization. Auerbach (1991) and Bradford (1995) discuss capital-gains tax systems that have the same incentive effects as accrual-based systems, but that tax gains when realized. These systems have not yet been tried in any practical context. In the United States, realized capital gains have frequently been taxed at different rates depending on their holding period. Because one of the policy objectives of those who argue for reduced tax rates on capital gains is to encourage long-term holding of securities, long-term gains have sometimes been taxed at a lower rate than short-term gains. Poterba and Weisbenner (2001b) present summary information on the changes over time in the US tax code's definition of "short-term" gains, which has varied between six months and one year.

The US tax system also limits claims for tax relief when security values decline and investors incur capital losses. These limits, which are known as loss-offset provisions, raise the effective tax burden on capital investments by making the tax rate at which gains are taxed when the asset appreciates higher than the tax rate at which losses can be deducted when the asset depreciates. Since losses are measured relative to the *nominal* historical basis in an asset, the value of loss offsets is reduced still further.

In addition to taxing capital income, some nations also tax wealth, although these taxes do not account for a substantial share of revenues in most developed nations. The United States does not have a wealth tax, but like many other nations, it does have an estate and gift tax that accounts for a nontrivial revenue share. Estate tax is levied on the total value of a decedent's estate plus the value of his taxable lifetime gifts. Gifts of up to $10 000 per recipient per year are excluded from the tax base. In 2000, estates valued at less than $675 000 were not taxed, but taxable estates valued at more than this amount face marginal estate tax rates that range from 35 to 60 percent. Poterba (2000b) reports data from the US Treasury Department showing that approximately 1.5 percent of deaths currently results in taxable estates. The nominal threshold for an estate to be subject to the estate tax is currently scheduled to rise to $1 million

by 2006, but there is ongoing legislative debate about the structure of the estate tax and the threshold at which estates become taxable.

In addition to the estate tax, most localities in the United States levy property taxes on real property. These taxes raise the effective tax burden on residential and non-residential land and on tangible assets such as equipment and structures. Some jurisdictions also tax consumer durables such as automobiles under a personal property tax. Financial assets are usually not included in property tax bases.

One of the most important developments in the tax treatment of capital income in the United States during the last two decades has been the growth of various ways to hold assets in "tax deferred" accounts. The Individual Retirement Account, which was effectively introduced in the Economic Recovery Tax Act of 1981, and was substantially limited by the 1986 Tax Reform Act, and the 401(k) retirement saving plan, are examples of such tax-deferred accounts. At the beginning of 2000, crude estimates suggest that the total market value of assets in 401(k)-type retirement saving plans exceeds $1 trillion, while the assets in Individual Retirement Accounts are at least twice as large. These assets represent more than five percent of household-net-worth, but for many middle-income households, they represent a much larger share. When IRA and 401(k)-plan assets are added to the assets in traditional corporate pension plans, it becomes clear that a substantial fraction of household financial assets are held in forms that do not generate current tax liability on capital income. Moreover, current trends suggest continued growth of assets in these tax-deferred accounts.

The United States is not alone in its use of tax-deferred accounts. In the United Kingdom, Personal Equity Plans provide a similar opportunity to accumulate assets and to pay tax only when income is drawn out of the account. Canadians can save on a tax-deferred basis through Registered Retirement Saving Plans. Poterba (2001a) provides an overview of the current limits on tax-deferred saving in a sample of OECD nations.

The foregoing discussion has ignored one of the most important capital-income taxes in most nations: the corporate income tax. While it is impossible to discuss some issues in portfolio choice, such as debt and equity clienteles, without reference to the corporate income tax, this tax is not the focus of the present chapter. Auerbach (2001) addresses many of the issues associated with the corporation tax in detail. While the corporate income tax affects the pre-tax returns that investors can earn on various assets, this chapter largely treats these pre-tax returns as given, and considers how households choose their portfolios in light of these returns.

1.2. The recent evolution of marginal tax rates on capital income in the United States

This section summarizes the recent evolution of capital income taxes in the United States. The 1980s and 1990s have been periods of unusual change in the structure of taxation in the United States, and the reforms over this period illustrate a range of different potential tax policies.

Prior to the Economic Recovery Tax Act of 1981 (ERTA), marginal tax rates on interest and dividends ranged up to 70 percent. Short-term capital gains were taxed as ordinary income, which meant that they could also be taxed at rates of up to 70 percent. Long-term capital gains, which were defined as gains on assets held for more than a year, were taxed at 40 percent of the ordinary income tax rate, which meant a top statutory rate of 28 percent. ERTA reduced the top statutory tax rate on interest and dividends to 50 percent, and therefore reduced the top rate on long-term capital gains to 40 percent.

Marginal tax rates on interest and dividend income were reduced still further by the Tax Reform Act of 1986 (TRA86), which reduced the marginal tax rate on highest-income individuals to 28 percent. Various tax changes since the 1986 reform have raised the top statutory tax rate from 28 percent in 1986 to 39.6 percent in 2000.

The tax treatment of capital gains has also changed significantly. TRA86 eliminated the preferential tax treatment of long-term gains, so that the statutory rate on long-term gains rose from 20 percent to 28 percent. For several years in the late 1980s, the highest statutory rate on realized capital gains was 33 percent, but this rate applied to taxpayers below the highest income levels. The capital-gains tax rate on the highest-income taxpayers remained at 28 percent until 1997, when the Taxpayer Relief Act of 1997 (TRA97) reduced the rate on long-term gains to 20 percent. In the year following the enactment of TRA97, there was an "intermediate-term" gain category that was subject to a tax rate between the rate on ordinary income (short-term gains) and 20 percent (long-term gains), but this intermediate classification was eliminated in 1999. Current legislation calls for a further decline, starting in 2005, in the statutory tax rate on very-long-term capital gains (gains on assets held for more than five years). The very-long-term capital-gains tax rate is scheduled to fall to 18 percent.

Table 1 provides summary information on the weighted average marginal income tax rate that applied to various types of capital income over the period 1979–1999. The results through 1995 are based on actual tax return data provided by the Statistics of Income Division of the Internal Revenue Service; the results for 1996–1999 are based on extrapolation from 1995 tax returns. These marginal tax rates are estimated using the NBER TAXSIM program, which is described by Feenberg and Coutts (1993). TAXSIM combines a detailed computer program for calculating individual tax liabilities with a database of individual income tax returns, released without individual identifiers, to summarize various aspects of the US income tax system. The first column shows the weighted average marginal income tax rate on dividend income, τ_d, which is defined as

$$\tau_d = \frac{\sum_{i=1}^{H} \tau_{\text{div},i} * \text{DIVS}_i}{\sum_{i=1}^{H} \text{DIVS}_i}.$$

Similar weighted average marginal tax rates for interest income and for realized capital gains are shown in columns two and three. For comparative purposes, the last column shows the weighted average federal marginal income tax rate on wage income.

Table 1
Individual marginal tax rates on capital income in the United States, 1979–1999[a]

Year	Dividends	Interest	Realized long-term gains	Wages	Tax-exempt interest
1979	28.8	41.7	16.5	27.6	n.a.
1980	30.5	42.8	16.8	29.1	n.a.
1981	31.1	40.8	17.1	30.2	n.a.
1982	28.2	35.4	15.1	28.4	n.a.
1983	25.7	33.6	15.2	26.6	n.a.
1984	26.6	33.0	15.2	26.2	n.a.
1985	26.6	32.8	15.5	26.4	n.a.
1986	25.8	32.6	16.2	26.6	n.a.
1987	24.2	27.9	25.3	24.2	29.1
1988	22.2	25.1	26.4	22.6	25.7
1989	22.9	25.4	25.9	22.7	26.0
1990	22.8	25.0	25.5	22.5	25.8
1991	22.6	25.5	24.4	22.6	25.9
1992	22.1	25.2	25.2	22.5	25.1
1993	23.6	27.2	26.0	23.3	27.9
1994	24.3	272	26.7	23.6	27.6
1995	25.3	28.2	26.9	23.9	28.9
1996	25.9	28.8	27.9	24.1	29.4
1997	26.5	29.8	24.6	24.5	30.5
1998	26.0	29.4	20.4	25.2	29.6
1999	26.6	29.7	20.5	25.4	29.8

[a] Source: NBER TAXSIM model calculations. Each entry presents a dollar-weighted average marginal tax rate on positive income amounts only, using data from the IRS Individual Tax Model, as analyzed with the NBER TAXSIM Model.

Table 1 illustrates the impact of recent tax changes on the relative tax burdens on different assets. The weighted average marginal tax rate on dividend income declined by five percentage points between 1980 and 1983, and by another 4.4 percentage points between 1985 and 1988. These changes were almost exclusively due to legislative changes. The second column of table 1 shows an even sharper decline in the weighted average tax rate on interest income between the late 1970s and the mid-1980s; this reflects difference in the distribution of interest and dividend income across income classes. The weighted average marginal tax rates on both interest and dividend income rise by several percentage points during the 1990s, primarily as a result of the increase in top marginal tax rates that was enacted in 1993.

The tax rate on realized long-term capital gains increased by 9.1 percentage points between 1986 and 1987 as a result of the Tax Reform Act of 1986, and the tax rates on realized gains in the late 1980s and early 1990s were substantially above the rates of the late 1970s. Between 1996 and 1999, the average statutory tax rate applying to long-term gains fell by 7.4 percentage points. This suggests that most realized long-term gains are taxed at the highest statutory rate, which declined from 28 to 20 percent.

The last column in table 1 shows the weighted average "implied" marginal tax rate on interest from tax-exempt bonds. This tax rate is higher than the weighted average tax rate on taxable interest income, although in some years by only a few tenths of a percentage point. This nevertheless suggests that the households who own tax-exempt bonds are in higher marginal tax brackets than those who own taxable bonds.

Weighted average marginal tax rates like those in table 1 provide some information on the incentive effects of tax policy, but they do not capture the substantial heterogeneity across households in the tax treatment of capital income. These differences play a critical role in determining which households will hold particular types of assets.

1.3. Household financial assets in the United States

Table 2 presents information on the relative importance of the various financial assets currently held by households in the United States. At the beginning of the year 2000, total household financial assets were valued at $35.6 trillion. Households also held tangible assets, primarily real estate and consumer durables, worth roughly one third as much as their financial assets. Within the set of financial assets, corporate stock accounted for $8.3 trillion, or approximately one quarter of the total. Mutual funds, which invest in equities more than other assets, account for another $3.2 trillion of household financial assets. Together, directly held stock and mutual funds comprise roughly one third of household financial assets. Equity in non-corporate businesses, which is relatively illiquid and is not usually traded in an organized market, is also substantial: it represents $4.6 trillion, or 13 percent of all financial assets. Another $5.9 trillion was held in taxable-interest-bearing instruments such as taxable corporate bonds, saving accounts, or Treasury bills. Table 2 illustrates the importance of tax-deferred asset accumulation: pension fund reserves account for more than $10 trillion, or between one quarter and one third of financial assets, in early 2000.

The bottom rows in table 2 show the value of net financial assets, subtracting either non-mortgage debt or all debt from the stock of financial assets. Roughly two thirds of the household liabilities shown in table 2 are home mortgages; the remainder is largely consumer credit. While much of the analysis in this chapter will focus on the allocation of household saving, it is important to consider how tax incentives affect borrowing behavior as well.

Table 2 shows the aggregate structure of the household balance sheet, but it does not capture the important cross-sectional heterogeneity in household asset holdings. The Survey of Consumer Finances is a rich data base on household assets and liabilities that

Table 2
Financial assets of US households, 2000[a]

Asset category	Applicable tax rate	Tax deferral?	Value (percent) of holdings
Deposits and money market funds	τ_b	No	4499.1 (12.6%)
Taxable bonds	τ_b	No	1428.1 (4.0%)
Tax-exempt bonds	Untaxed	Not applicable	535.3 (1.5%)
Corporate equity	τ_{div}, τ_{cg}	Yes	8266.7 (23.2%)
Mutual fund shares	τ_{div}, τ_{cg}	Some	3186.3 (9.0%)
Life insurance reserves	$\tau_{ordinary}$	Yes	791.6 (2.2%)
Pension reserves	$\tau_{ordinary}$	Yes	10395.6 (29.2%)
Personal trust investments	$\tau_{ordinary}$	Yes	1135.2 (3.2%)
Equity in unincorporated businesses	τ_{cg} ??	Yes	4639.6 (13.0%)
Miscellaneous assets	Varied	Possibly	708.0 (2.0%)
Total financial assets			35585.7
Home mortgages	$\tau_{ordinary}$	No	(4547.6)
Other debt	$\tau_{ordinary}$	No	(2420.7)
Net financial assets			28617.4
Financial assets net of non-mortgage debt			33165.0

[a] Source: Author's calculations based on Federal Reserve Board, Flow of Funds Accounts of the United States: Flows and Outstandings, First Quarter 2000 (Release Z.1). Calculations are based on reported information for the household sector, which includes nonprofit institutions. In 1996, the last year for which detailed information on the portfolio holdings of nonprofit institutions are available, "non-nonprofit holdings" represented 5.7% of total household holdings of financial assets. Values in parentheses are percentages of total financial assets. For most households, $\tau_{div} = \tau_b = \tau_{ordinary}$, where τ_{div} is the tax rate on dividends, τ_b the tax rate on interest, and $\tau_{ordinary}$ the tax rate on ordinary income.

provides such disaggregate information every third year for households in the United States. Kennickell, Starr-McCluer and Surette (2000) and Bertaut and Starr-McCluer (2001) report on the most recent patterns of asset holding as reported in the 1998 wave of the survey. These data also form the basis for a number of the empirical studies discussed below.

There is less empirical evidence on the aggregate structure of household portfolios for nations other than the United States than for the United States. Information for several developed nations, based on household-level surveys, is collected in Guiso, Haliassos and Jappelli (2001).

2. Taxation and portfolio structure

The central question in the analysis of taxation and portfolio structure is how tax-induced distortions in after-tax returns affect investors' asset demands. Poterba

Table 3
Returns on portfolio assets, 1926–1996[a]

Asset/return concept	Return (%)	Standard deviation (%)
Part A: Pretax returns		
Pretax nominal returns:		
Large stocks	12.67	20.32
LT Government bonds	5.45	9.21
Treasury bills	3.78	3.26
Pretax real returns:		
Large stocks	9.45	20.91
LT Government bonds	2.23	10.85
Treasury bills	0.57	4.36
Part B: After-tax returns		
After-tax nominal returns:		
Large stocks	9.16	17.15
LT Government bonds	3.39	7.03
Treasury bills	2.15	1.69
After-tax real returns:		
Large stocks	5.94	17.89
LT Government bonds	0.17	9.25
Treasury bills	−1.07	4.38

[a] Source: Author's calculations using pretax return data, and information on the decomposition of returns into income and capital gains, reported in Ibbotson Associates (1996). Marginal tax rates for the "after-tax" calculation correspond to tax rates on a joint filer with a constant $1989 income of $75 000; this marginal tax rate is drawn from Siegel and Montgomery (1995).

(2001a) explains that one can identify such distortions along six different margins: asset selection, asset allocation, borrowing, asset location in taxable and tax-deferred accounts, asset turnover, and the choice of whether or not to hold assets through various financial intermediaries.

This section considers the link between taxation and asset choice. It begins with a summary of the theoretical models that have been proposed for analyzing taxation and portfolio choice, and then examines the available empirical evidence on the impact of taxation on the structure of household portfolios.

Before considering specific models, it is important to recognize the significant effect that taxation can have on the set of returns available to investors. Table 3 presents summary information on the vector of returns on three asset classes, large stocks, long-term government bonds, and Treasury bills, that were available to US investors

over the seven-decade period 1926–1996. The first panel shows the average before-tax nominal and real returns available to an investor. This panel indicates the returns available to an untaxed investor, such as a nonprofit institution. The lower panel shows the set of after-tax returns available to an investor with an income of $75 000 in 1989, under the assumption that this individual's real income remained the same in all years. The real after-tax return on equities falls from 9.5 percent to 5.9 percent, or by nearly forty percent. The real after-tax return on long-term bonds falls from 2.2 percent to 0.2 percent, and for Treasury bills, the real after-tax return averages −1.1 percent. The return differentials would be even larger if the taxpayer was assumed to have a higher income and therefore to face higher marginal tax rates. Ghee and Reichenstein (1996) present further discussion of the difference between pre-tax and after-tax returns, and the importance of such differences for investor behavior.

The set of investment decisions that an investor would make would depend on whether he confronted the pre-tax returns or after-tax returns. Not only are the returns on all assets substantially lower on an after-tax basis, but also the relative returns on different assets are different. Equity, which generates a substantial fraction of its returns in the form of lightly taxed capital gains, becomes relatively more attractive when returns are measured on an after-tax basis than a before-tax basis.

2.1. Asset demand in clientele models

To develop some insight in the effect of taxation on portfolio structure, it is helpful to begin with simple models in which two or more assets yield the same pre-tax returns in all states of nature. If the tax rules governing the returns on these assets are the same for all investors, and if the tax treatments of the two assets are different, then portfolio equilibrium requires that the prices of the assets adjust so that the expected after-tax returns on the two assets are identical. In this case, each investor will be indifferent between holding the two assets.

When different investors are taxed in different ways on the two assets, the analysis becomes more complex. In this case, investor clienteles will emerge in the holdings of various securities. The simplest and best-known clientele model is Miller's (1977) model of the choice between debt and equity. In his framework, debt and equity are both riskless, so investors decide which security to hold only on the basis of after-tax returns. Miller assumes that equity returns are untaxed to all investors, but that investors are taxed on interest income, and that their interest-income tax rates vary. The result is a clientele equilibrium in which high-tax-bracket investors hold corporate equities, and those in lower tax brackets hold corporate debt. For a given set of pre-tax returns on equity and debt, r_{eq} and r_b, the asset demand functions (E^d and B^d) for an investor with net worth W can be written as

$$E^d = W,\ B^d = 0 \quad \text{if} \quad (1 - \tau_b) r_b < r_{eq},$$
$$E^d = 0,\ B^d = W \quad \text{if} \quad (1 - \tau_b) r_b > r_{eq}.$$

This model predicts that investors will hold completely specialized portfolios.

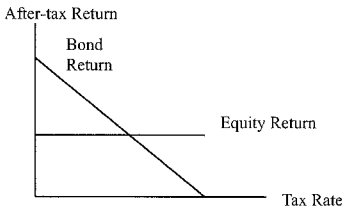


Fig. 1. Equilibrium in the Miller model.

Market equilibrium is determined by combining the asset demand conditions given above with asset supply functions that come from firm decisions designed to minimize the cost of funds. Because interest payments can be deducted from corporate income taxes, the after-tax cost of debt finance is $(1-\tau_{corp})*r_b$. The after-tax cost of equity finance is simply r_{eq}. Thus, if $(1-\tau_{corp})*r_b < r_{eq}$, firms will supply debt, while if the opposite inequality holds, they will supply equity. To avoid excess supply of either debt or equity, equilibrium requires $(1-\tau_{corp})*r_b = r_{eq}$.

The relative pre-tax returns on debt and equity determine which investors will invest in debt and which will invest in equity. Figure 1 illustrates the asset ownership clienteles that form in this model: any investor for whom $(1-\tau_b)/(1-\tau_{corp}) < 1$ will hold equity, and anyone for whom the inequality is reversed will hold debt. The "marginal investor", the investor who is indifferent between debt and equity, has a tax rate on interest income equal to the corporate tax rate: $(1-\tau_b)/(1-\tau_{corp}) = 1$. For this investor, the after-tax return on bonds, $(1-\tau_b)*r_b$, just equals the after-tax return on equity, $(1-\tau_{corp})*r_b$.

The Miller (1977) model provides a useful illustration of how asset market clienteles could emerge, and how asset demands can be combined with asset supply conditions to determine the equilibrium returns on different securities. The discussion here focuses primarily on the Miller model's clientele equilibrium structure, not its implications for corporate finance. But it is worth noting that while Miller (1977) suggests this model as a description of the debt–equity behavior of corporations in the United States, the model does not appear to perform well on this front. The model predicts that relatively few households will demand equity rather than debt, and for some configurations of tax parameters that have been observed, it predicts the complete absence of corporate equity.

This difficulty can be illustrated using tax rates in the United States in tax year 2000. The current corporate income tax rate is 35 percent. Assuming, counterfactually, that equity returns are untaxed at the investor level, the only households who should hold equity are those facing marginal tax rates on interest income above 35 percent. If equity returns are taxed, as the amount of dividends and equity capital gains reported on tax returns suggests they are, then the marginal tax rate on interest income at which investors will be indifferent between debt and equity securities is higher than 35 percent. In 2000, with the top marginal personal tax rate above 34 percent, the Miller model suggests that the only taxpayers who should hold equity are those in the

top two marginal-income tax brackets (36 percent and 39.6 percent). Equity is actually held by taxpayers with marginal tax rates well below 34 percent.

In the late 1980s and early 1990s, the Miller model's difficulty in explaining observed debt–equity ratios was even more acute. The top individual income tax rate was either 28 or 33 percent, and the corporate income tax rate was 34 percent. In this setting, the simple Miller model would have predicted the complete absence of equity in the US economy! Despite the difficulties with the simple Miller model, the basic clientele insight can provide a starting point for a richer analysis of portfolio decisions. Mintz and Smart (2001) illustrate such an approach, in which investors with tax losses and traditional taxable investors interact to determine portfolio clienteles. Clientele models like the one that Miller (1977) used to study debt policy can also be applied to the analysis of dividend policy; Allen, Bernardo and Welch (2000) provide a recent example.

2.2. Taxation and asset demands with risky returns

Before considering the effect of taxing many risky assets, it is helpful to develop some intuition using a simple framework, pioneered by Domar and Musgrave (1944), with one risky asset, and one riskless asset. This analysis draws attention to the important distinction between private and social risk-taking, and the potential effect of tax policy on the fraction of society's assets that are invested in risky securities.

Domar and Musgrave (1944) showed that when taxes are levied on the excess return from a risky asset, and when gains are taxed and losses are deducted without limit at the same rate, then the government effectively becomes the investor's partner. They argued that this could lead private risk-takers to increase their total assets at risk, since the tax would lead to equiproportional reductions in the expected return and the risk of potential projects. To consider this case, let r_i denote the return on a risky asset, and r_f the return on a riskless asset. The investor's after-tax return on the risky asset is

$$r_{i,at} = r_f + (1 - \tau)*(r_i - r_f).$$

After-tax wealth at the end of one period is

$$W_{at} = (1 - a)*r_f*W + a*[r_f + (1 - \tau)(r_i - r_f)]*W = r_f*W + a*(1 - \tau)*(r_i - r_f)*W.$$

Notice that in this expression, the term $(1 - \tau)$ always enters as a product with a, the fraction of the investor's net worth that is held in the risky asset. Thus if τ rises and $(1 - \tau)$ declines, the investor can preserve the same after-tax opportunity set as before the tax by increasing a. If a (a') denotes the amount that the individual would have invested in the risky asset in a no-tax (taxable) world, then provided $a' = a/(1 - \tau)$, the individual's after-tax wealth is unaffected by the presence of the tax. The individual bears the same level of risk by investing a' in the risky asset in a taxable world and

a in a taxless world, but the social level of risk-taking is greater in the taxable case, since $a' > a$.

The Domar–Musgrave analysis does not speak to the issue of how the risk of uncertain tax collections is allocated across individuals, and in particular, how risk is allocated through the tax system. Subsequent research, notably by Bulow and Summers (1984), Gordon (1985), and most recently Kaplow (1994), has embedded the Domar–Musgrave analysis in a general equilibrium setting, and recognized the effect of risky tax receipts on the government budget constraint. When individual investors are indifferent to incremental investments in the risky projects that face taxation, and the government is no more efficient than private capital markets at spreading risk through the economy, this literature shows that a proportional tax claim on all gains and losses has no market value.

Another special feature of the Domar–Musgrave analysis is its focus on taxes on excess returns. Actual tax systems usually tax total returns rather than excess returns. While taxes on excess returns have substitution effects, but still permit the investor to achieve the same riskless return, taxes on total returns affect both the relative returns on different assets and the overall *level* of returns. As such they have both substitution and wealth effects. When the tax rate that applies to the total return on asset i rises, the substitution effect leads investors to demand less of this asset. In addition, however, as a result of the reduction in the after-tax return to asset i, there is a wealth effect of ambiguous sign. If the wealth elasticity of demand for risky assets is positive, then an increase in the tax rate on asset i will reduce the amount held in this asset as the substitution and wealth effects will work in the same direction. This result is derived in Sandmo's (1985) survey, which draws on Stiglitz (1969) and the generalization to multiple assets in Sandmo (1977).

One important extension of the Domar–Musgrave analysis is to the case of imperfect loss-offsets. The practical justification for limiting the losses that investors may claim is that without such limits, firms or individuals could undertake projects that generate private benefits but taxable losses, and thereby collect government subsidies for what are effectively private consumption activities. Loss-offset provisions make such transactions more difficult, but at the cost of raising the effective tax burden on legitimate projects that face some risk of generating losses. Previous work on loss-offsets, such as MacKie-Mason (1990), considers how tax systems with limited loss-offsets affect the certainty equivalent present discounted value of the after-tax project returns. The general equilibrium effect of imperfect loss-offset, and the effect of such provisions on the required returns for risky assets, has not been analyzed.

2.3. The after-tax capital asset pricing model and asset demands

Actual asset markets offer investors a wide range of risky securities which generate income streams that are taxed at different rates. The pre-tax returns on these assets are imperfectly correlated, so the actual portfolio problem confronting households involves choosing assets on the basis of both their tax and risk characteristics. Asset demands in

this setting have been analyzed by Auerbach and King (1983), Brennan (1970), Elton and Gruber (1978), Litzenberger and Ramaswamy (1979), Long (1977) and Talmor (1985). To describe the structure of asset demands, it is necessary to develop some notation. Let W_0 denote a household's beginning-of-period investable wealth, and E_i denote the household's investment in risky asset i. Assume that the riskless rate of return is r_f, that this asset pays all of its return in the form of interest, and that interest income is taxed at rate τ_b. Assume that all risky assets are taxed at a rate of τ_{eq} (generalizing to the case of different tax rates on different risky securities is tedious, but straightforward), and that pre-tax returns on equity securities are given by r_i, where $i = 1, \ldots, N$. The expected pre-tax return on equity security i is μ_i. Denote the vector of mean returns on equity securities $\{r_1, \ldots, r_N\}$ by μ, and assume that Σ denotes the N-by-N covariance matrix of risky returns.

The individual investor is assumed to maximize a utility function that can be written in terms of the mean and variance of final wealth, $U(W, \sigma_W^2)$. Using the foregoing notation to define expected end-of-period wealth, and its variance, as a function of the return generating parameters and the amounts invested in each equity security, and substituting into the utility function, yields the function to be optimized:

$$U\left(\left[W_0 - \sum_i E_i\right] * (1 - \tau_b) r_b + \sum_i E_i * (1 - \tau_{eq}) r_i, \sum_i \sum_j E_i E_j * (1 - \tau_{eq})^2 * \sigma_{ij}\right).$$

The first-order condition for optimal holdings of risky asset i is given by:

$$U_W * \left[-(1 - \tau_b) r_b + (1 - \tau_{eq}) r_i\right] + 2 * U_{\sigma^2} * (1 - \tau_{eq})^2 * \sum_j E_j * \sigma_{ij} = 0.$$

If we define

$$\delta = \frac{U_W}{2 * U_{\sigma^2} * (1 - \tau_{eq})^2},$$

then the first-order condition for optimal asset holding can be rewritten as

$$\delta * \left[-(1 - \tau_b) r_b + (1 - \tau_{eq}) r_i\right] = \sum_j E_j * \sigma_{ij}.$$

This expression can be rewritten in matrix notion, using $E = \{E_1, \ldots, E_N\}$ as a column vector and **1** as a column vector of ones, as:

$$\delta * (1 - \tau_{eq}) \mu - \delta * (1 - \tau_b) r_b * \mathbf{1} = \Sigma^* E.$$

The resulting set of optimal holdings of the risky assets then satisfies

$$E^* = \delta * \Sigma^{-1} * \left[(1 - \tau_{eq}) \mu - (1 - \tau_b) r_b * \mathbf{1}\right].$$

In the special case of no taxes on interest income or equity returns, this expression reduces to the standard asset demand expression, $E^* = \lambda^* \Sigma^{-1} (\mu - r_b * \mathbf{1})$, where

$\lambda = U_W/2U_{\sigma^2}$. When returns are taxed, Auerbach and King (1983) show that the optimal portfolio can be interpreted as a weighted average of two portfolios, one of which is the market portfolio, and the other of which is a portfolio that is chosen on the basis of tax but not risk considerations. The relative weights on these two basic portfolios depend on the investor's tax rates in comparison to the tax rates of other investors, and on the investor's degree of risk aversion. A more risk-averse investor will place greater weight on the portfolio that is efficient from the perspective of diversification, and correspondingly down-weight the portfolio that derives from tax specialization.

While several studies have explored the structure of asset demands in the presence of heterogeneous investor taxation, the general equilibrium structure of asset markets in this setting has received much less attention. The essential problem is that differential tax rates on different investors present opportunities for tax-motivated arbitrage. Unless there are limits on the size of the net positions that investors can hold in various assets, asset-market equilibrium may not exist. A simple example illustrates this point. Consider an economy with two risk-neutral investors. One is an untaxed institution, and the other is a taxable investor who is not taxed on the income from equities, but who is taxed at rate τ on interest income and is permitted to deduct interest payments. If there are no restrictions on long or short sales, then the taxable investor will borrow an infinite amount from the untaxed investor, and pay tax-deductible interest at a rate r. The untaxed investor will offset these transactions by issuing corporate stock, which will be purchased by the taxable investor. Since neither investor is taxed on equity income, the two investors will collect $\tau * r$ dollars from the government, in the form of a tax rebate for interest deductions, for every dollar of debt and equity that is issued. Unless a constraint prevents this tax arbitrage, it will continue without limit.

Several studies, including Ross (1985), Dammon and Green (1987) and Basak and Croitoru (2001), have derived strong conditions under which such tax arbitrage will not take place and explored the nature of the resulting market equilibrium. These results serve primarily to underscore the difficulty of achieving equilibrium in a capital market without transaction costs but with heterogeneous investor taxes. Auerbach and King (1982, 1983) explore the effect of short-selling constraints on the asset-market equilibrium, and address related issues such as whether or not investors in a firm will agree on the firm's optimal debt–equity mix. Further work remains to be done both on the nature of equilibrium in asset markets with plausible imperfections, such as short-selling limits or transactions costs, and on the welfare effects of tax policy in such markets. Basak and Gallmeyer (1998) is a recent study that addresses some of these issues.

There is growing recognition that the asset allocation problem in the presence of taxes is much more complex than the analogous problem without taxes. For example, Meehan, Yoo and Fong (1995) resort to numerical solutions to evaluate the after-tax asset allocation problem. The problem of asset selection in the presence of taxes is currently attracting growing attention both from academics and from practitioners interested in delivering advice to high-net-worth taxable clients.

2.4. Empirical evidence on taxation and portfolio choice

A number of empirical studies have investigated the links between the structure of household portfolios and the taxation of capital income. These studies broadly suggest that taxes do affect asset-ownership patterns, in contrast to much earlier surveys, such as Butters, Thompson and Bollinger (1953) and Barlow, Brazer and Morgan (1966), that conclude that taxes have little effect on the portfolio decisions of high-net-worth households. A range of data and statistical difficulties, however, besets many of these empirical studies. It is important to identify two of these issues at the outset.

The first is that there are very few datasets that include any information on the high-net-worth households whose behavior is central to studies of taxation and portfolio behavior. Most household surveys that are based on random population samples have very low response rates among high-income, high-net-worth households. Moreover, these surveys rarely collect sufficiently disaggregated information on asset holdings to permit the type of data analysis that is required to test tax-based theories of portfolio choice. For example, tax-oriented theories suggest that it is important to distinguish between corporate stock held through mutual funds and shares held directly, because these alternative means of holding equities have different tax consequences. It is also important to distinguish between mutual funds invested primarily in corporate stocks and those that hold government and corporate bonds. Yet most surveys that inquire about asset holdings, if they ask about assets at all, group together stocks and mutual funds, and do not inquire about the types of mutual funds held.

The second important difficulty in developing empirical tests of portfolio behavior is conceptual. The set of asset classes that investors may choose to invest in is large. It includes corporate stock, mutual funds invested in stocks or bonds, taxable bonds (government or corporate), short-term interest bearing accounts such as saving accounts, CDs, and money-market accounts, tax-exempt bonds, investments in venture capital startup firms or similar partnership ventures, owner-occupied real estate, commercial real estate, and international securities (stocks or bonds). In addition, with the exception of owner-occupied housing, any of these assets could be held directly, in a taxable form, or in a tax-deferred form such as through a defined contribution pension plan or an Individual Retirement Account.

Almost no households hold assets in each of the broad categories described above. This raises questions about the value of theories that emphasize that all investors should hold the market portfolio, and it also raises an empirical problem for studies of how taxes affect the portfolio shares allocated to different assets. Asset demands must be modeled conditional on the set of assets that a household owns, and this requires first-stage modeling of why investors hold incomplete portfolios.

A small literature has focused on the extent of portfolio incompleteness and tried to explain this phenomenon. King and Leape (1998) document the incompleteness of household portfolios using data from a special high-net-worth survey conducted by the Stanford Research Institute in 1978. Leape (1987) discusses potential explanations for this incompleteness, emphasizing the information costs of learning about different

assets. Haliassos and Bertaut (1995) have recently addressed the incompleteness issue from a different perspective, asking in particular why so few households own corporate stock. They rule out explanations involving minimum investment requirements and risk aversion, and argue that inertia and departures from expected utility maximization provide more promising explanations for the observation that, in the 1980s, nearly three quarters of US households did not hold corporate stocks.

While the explanation of the "non-stock-ownership" puzzle is important, the empirical magnitude of the puzzle has declined in the last decade. Data from the 1998 Survey of Consumer Finances, reported in Kennickell, Starr-McCluer and Surette (2000), as well as the Investment Company Institute (1999), suggest that roughly half of the households in the United States currently own some corporate stock.

The endogeneity of the set of assets held makes it essential to treat asset-demand decisions as a two-step process. The first involves the decision of what assets to hold, and the second concerns the decision of how much to hold in each asset class. The endogeneity of the set of assets with positive holdings is a difficult empirical problem, and one that has been addressed in a variety of *ad hoc* ways in previous empirical studies. These approaches may in part explain why it has proven easier to identify important effects of taxation on the set of assets held than on the level of assets held in different forms, conditional on ownership.

Feldstein (1976) presented the first systematic econometric analysis of taxation and portfolio choice. He used data from the 1962 Survey of Financial Characteristics of Households, which was a precursor to the Survey of Consumer Finances. He studied the probability that a household owned corporate stock, which is a tax-favored asset because at least part of the returns are earned in the form of capital gains which are taxed less heavily than interest and dividend income. He found that higher current income was associated with a higher probability of equity ownership, conditional on household net worth, and he therefore concluded that higher marginal tax rates, which are associated with higher income, discourage equity ownership. He also presented results for several other asset categories, including taxable and tax-exempt bonds. His results for corporate equities suggested that a ten-percentage-point increase in the marginal tax rate on interest and dividend income could lead to a 3.7-percentage-point increase in the probability of equity ownership. Because the maximum federal marginal tax rate was 91 percent at the time of the survey, the potential for substantial tax-induced distortions to portfolio behavior was large.

Feldstein's (1976) study, while pioneering, left several issues unresolved. Most importantly, the fact that all of the identification for marginal tax rate variation was generated by income differences raises issues of interpretation. While wealth rather than income is the traditional argument in asset-demand models, and his estimating equations included household net worth, income might still be correlated with asset ownership for non-tax reasons. If higher-income households have different asset demands than lower-income households with the same net worth, as they might if high current income indicates high future income and therefore different amounts of human capital, then the observed differences in asset holdings could

be due to the correlation between human capital, asset holdings, and current income.

Feldstein's (1976) analysis also ignored the statistical problems that arise when substantial numbers of households report no holdings of major asset classes. The subsequent study that has addressed this problem most carefully is King and Leape (1998). This study models the set of discrete choices that are associated with the decisions to hold assets in particular classes. It uses the results of this discrete-choice analysis to correct for the econometric biases that could result from estimating asset-demand equations without a selection correction for those households with positive holdings. The central empirical conclusion is that taxes have substantial effects on the set of assets held by different households, but they have relatively small effects on the portfolio shares conditional on ownership.

Hubbard (1985) provides further support for the Feldstein (1976) conclusion. He uses a unique data set collected by the US President's Commission on Pension Policy in 1979 and 1980. These data make it possible to construct a measure of a household's future pension income and Social Security benefits, and to control for these components of wealth in modeling asset demands. This work provides valuable information on how the liquid component of household financial assets is allocated.

The advent of the Survey of Consumer Finances database, which began in 1983 and has been extended every three years since then, has permitted a number of more recent studies of taxation and household portfolio structure. The major tax reforms of the 1980s have also created valuable opportunities for studying how taxation affects asset demands. Scholz (1994) studies the portfolio patterns in the 1983 and 1989 SCFs, and while he finds an important effect of TRA86 on the level of tax-deductible borrowing, he does not find any clear evidence of other portfolio shifts. One provision of the 1986 reform eliminated the deductibility of interest on non-mortgage debt, and this induced high-income taxpayers who itemize deductions on their tax returns to shift toward home-equity lines of credit, or mortgage indebtedness, and away from other types of consumer credit.

Scholz' (1994) finding of increased mortgage borrowing at high income levels has been confirmed, using other data sources, by Maki (1996). However, his conclusion that TRA86 did not affect the structure of household portfolios does not appear to be robust.

Poterba and Samwick (2002) use changes in tax rates as well as cross-sectional tax-rate heterogeneity to identify the effects of taxation on asset demands in the 1983, 1989, 1992, 1995, and 1998 Surveys of Consumer Finances. They do not address the endogeneity of asset holdings in as much detail as King and Leape (1998), but they estimate probit and tobit models for asset ownership. They include covariates similar to those in Feldstein (1976) and control for both income and wealth in asset-demand equations, although they, like other studies, encounter the problem of controlling for differences across households in risk tolerance that are not correlated with other observable variables. Poterba and Samwick (2002) find a substantial effect of a household's marginal tax rates on its probability of owning corporate stock,

tax-exempt bonds, or a tax-deferred account. Their results on the effect of taxation on the share of a portfolio held in different assets are weaker than the findings for ownership structure.

The explanation of disparities between Scholz (1994) and Poterba and Samwick (2002) is most likely in differential opportunities for investors to respond to the major tax reform of 1986. Scholz (1994) compares the 1983 and 1989 SCFs, and it is possible that households take time to modify their portfolios, so that few differences were observable only three years after the tax reform took effect. Other researchers have also noted some anomalies in the 1989 Survey of Consumer Finances, such as a decline in the total value of corporate equity held by households between the 1983 and 1989 SCFs, despite the rapid rise in share prices over this six-year period. Data anomalies in the 1989 SCF could also contribute to explaining the difference in results.

Samwick (2000) has also examined the impact of taxation on portfolio structure, focusing on time-series changes in household asset ownership. He chronicles the set of tax changes in the United States during the last two decades, and concludes that taxes can explain only a small fraction of the changes in household portfolio structure over this time period. One difficulty with focusing on time-series rather than cross-sectional changes, however, is that it becomes essential to consider general equilibrium effects associated with corporate as well as personal tax rules.

The lack of household-level data on portfolio structure in countries other than the United States has limited the scope of research on taxation and household portfolios. One notable study that does parallel the recent work in the United States is Agell and Edin's (1990) investigation of asset data reported on the annual Swedish income distribution survey. This study recognizes and treats the incomplete portfolio problem in the same fashion as King and Leape (1998), but it aggregates assets in order to facilitate estimation. The results support an important effect of taxation on portfolio choice. With respect to common stock, for example, a one-percentage-point increase in the marginal tax rate on interest income is predicted to increase the percentage of net worth allocated to equities by two percent, i.e. from 20 percent to 20.4 percent. The effect on participation in tax-advantaged saving schemes is even larger. Hochguertel, Alesie and van Soest (1997) and Stephens and Ward-Batts (2001) are other examples of studies that use non-U.S. data to explore taxation and portfolio issues.

The foregoing analysis has focused on how taxation affects the allocation of household financial assets, without considering the role of real assets in household portfolios. Ioannides (1989) and Berkovec and Fullerton (1992) estimate demand functions for owner-occupied real estate and other assets on the household balance sheet. The relationship between real and financial assets requires further investigation. It attracts particular attention with respect to elderly households, many of whom have accumulated substantial stocks of real assets, such as owner-occupied housing, but relatively small balances of financial assets. Tax-induced distortions in the demand for

real estate, particularly owner-occupied housing, can also play an important role in calculations of the welfare cost of the existing tax system.

2.5. *Taxation and investor clienteles for corporate stock: dividends vs. capital gains*

The empirical studies described above considered how the marginal tax rates faced by different households affected the probabilities that they held particular assets, and the fraction of their wealth that they allocated to different assets. A distinct literature has focused on the choice of assets *within* broad asset classes, and in particular the effect of marginal tax rates on investor's decisions to hold corporate stocks with high rather than low dividend yields. When capital gains are taxed at lower marginal tax rates that dividends, households who face high marginal tax rates on dividend and interest income have an incentive to hold more of their portfolio in stocks and to concentrate their portfolio in shares that generate capital gains rather than dividends.

The empirical literature on equity portfolio yields and marginal tax rates dates at least to Blume, Crockett and Friend (1974). Using a unique data file based on dividend income reported on tax returns from the late 1960s, and with information on the individual securities that investors held, they examined the relationship between a household's adjusted gross income and its portfolio dividend yield. The results suggested that households facing higher marginal tax rates, and therefore higher burdens on dividend income relative to capital gains, held portfolios with lower dividend yields, but the absence of control variables for household wealth or other household characteristics makes it difficult to evaluate their findings.

A number of subsequent studies have provided further evidence on the correlation between investor marginal tax rates and the dividend yield on common stock holdings. Both Petit (1977) and Lewellen, Stanley, Lease and Schlarbaum (1978) analyze the same data set on portfolio holdings and transactions by the clients at a major US brokerage house during the 1960s. They reach different conclusions, with the former providing support for the clientele hypothesis and the effect of taxes on portfolio yields more generally, and the latter finding relatively small effects of taxation on yield. Petit's (1977) results are more transparent, since they are based on regression methods, but the substantial differences in the study findings is puzzling and possibly attributable to small differences in the set of observations being analyzed. Chaplinsky and Seyhun (1990) also present evidence on clientele models, in their case relying on data from tax returns. They show that the ratio of dividends received to realized capital gains declines as household marginal tax rates on dividend income increase, but this finding does not necessarily support clienteles with respect to dividend yield. Numerous studies, as noted below, have found that capital gains realizations are sensitive to tax rates, and since realized gains are the denominator for the ratio that is used as the dependent variable, the tax sensitivity of realizations could be driving the results.

Scholz (1992), who analyzed Survey of Consumer Finances data from 1983, reports the most recent evidence on dividend clienteles. He controls for the constraint that

households cannot hold portfolios with negative yields, and finds a very pronounced effect of taxes on portfolio yields. He estimates a very large effect of tax rates on dividend yields. The difference between the predicted dividend yield assuming that the marginal tax rate is 50 percent, and the predicted yield assuming that there are no taxes, is 5.4 percentage points. This differential is substantially larger than the disparity in dividend yields between the highest-yield and lowest-yield deciles of traded equities, so there is some question of whether the estimated effect is implausibly large. One concern is that unobserved differences in risk tolerance might explain some of the findings.

The magnitude of the findings notwithstanding, there are strong reasons for preferring results from the Survey of Consumer Finances to those from all of the previous empirical studies in this area. The SCF has the important virtue of providing direct information on the market value of corporate equity holdings, so it is possible to calculate an actual dividend yield, and it also provides far more control variables relating to demographics and household characteristics than other studies using tax-return data. It is striking that some of the clearest evidence of dividend yield clienteles with respect to corporate stock ownership comes from a period when marginal tax rates on dividends, which were capped at 50 percent in 1983, were lower than when the other data sets were collected.

Most of the previous research on investor clienteles with respect to dividend yields has focused on the case of individuals, and there is much less evidence on how taxation affects the behavior of institutional investors. Strickland (1996) presents some informative evidence on this issue, and finds that taxable institutions such as mutual funds and money managers exhibit a preference for low-yield stocks, while untaxed institutions, such as pension funds, do not display any investment preference with respect to a firm's dividend yield. This represents further evidence that taxation affects the structure of investors' equity portfolios.

2.6. Asset market evidence on investor valuation of dividends and capital gains

The empirical research described above presents direct evidence on how taxation affects investor demand for dividends and capital gains, but it does not consider the effect of tax-induced shifts in investor demand on the market prices of securities that generate returns in the form of dividends and, alternatively, capital gains. Because the market prices of such securities represent key signals to corporate managers who are trying to determine their firms' financial policies, understanding how taxation affects the equilibrium market valuation of dividends and capital gains is a critical empirical task. Moreover, given the heterogeneity in the relative tax burdens on capital gains and dividend income for different investors, researchers have been interested in trying to identify the tax rates of the "marginal investor" who sets prices.

This task has attracted substantial research attention in both financial economics and public finance. The voluminous literature on this topic can be explained both by the central role of this issue in understanding corporate payout policy, as well as by

the accessibility of data and the straightforward nature of the empirical tests that are associated with it.

The test that one would like to perform is to compare the prices of two otherwise equivalent securities, one of which produces returns in the form of taxable dividends, the other generating the same pre-tax returns but in the form of capital gains. Such a comparison is typically not possible, because there are virtually no pairs of securities that generate returns that are taxed in different ways. One notable exception is the Citizens Utilities Company, a Connecticut utility firm, which has two classes of common stock that pay dividends that are taxed in different ways. Long (1978) and Poterba (1986a) analyzed the relative prices of these shares, and found mixed evidence on the impact of taxation.

A much larger literature focuses on the relative value of dividends and capital gains in dividend-paying firms. This literature, which involves two types of tests, is well surveyed by Allen and Michaely (1995). One set of tests asks whether the pre-tax return on corporate shares is systematically related to their dividend yields. This involves comparing ex post returns over periods when dividend-paying firms are paying dividends, and even when they are not, with the returns on comparable firms that do not pay dividends. In essence, the key question is whether a high-yield firm is required to earn a higher, or lower, return *at all times* than a lower-yield firm. Evidence on this issue is mixed. Litzenberger and Ramaswamy (1979, 1980) find that higher-yield securities generate higher pre-tax expected returns, which is consistent with the after-tax capital-asset pricing model. Naranjo, Nimalendran and Ryngaert (1998) also present evidence supporting such a relationship between yield and return, although they conclude that their empirical findings are too large to be accounted for by tax effects alone.

The second strand of literature, and the primary focus of the discussion below, asks whether the pre-tax return on dividend-paying firms differs from that on non-dividend firms *on the days when dividends are paid.* This is the "ex-dividend day" pricing literature, which compares the share-price decline on the day when investors are no longer eligible to receive a dividend with the amount of the dividend payment, and uses this ratio to estimate the relative tax burden on dividends and capital gains. One reason that ex-dividend-day pricing tests have received so much attention is that they offer a relatively straightforward test of the valuation of two income flows with different tax treatment.

The basic logic of the ex-dividend day testing strategy can be illustrated as follows. Consider a setting in which all investors face tax rates of τ_{div} and τ_{cg} on dividends and capital gains, respectively, and when their required after-tax return is ρ. In this case the equilibrium condition that ensures that all investors are indifferent to holding more or less of a security with a dividend yield of d_i and an expected capital gain of g_i is

$$\rho = (1 - \tau_{\text{div}})*d_i + (1 - \tau_{\text{cg}})*g_i.$$

This equation implies that cross-sectional variation in dividend yields should be reflected in differences in the total pre-tax return on different shares. In particular,

since the pre-tax return on a security is $R_i = d_i + g_i$, manipulation using the foregoing equilibrium condition reveals that

$$R_i = \frac{\rho}{1 - \tau_{cg}} + \frac{\tau_{div} - \tau_{cg}}{1 - \tau_{cg}} * d_i.$$

Thus, if dividend income is taxed more heavily than capital-gains income for the investors who determine market prices, higher-dividend-yield securities should earn higher pre-tax returns. Because there is substantial heterogeneity in the dates on which firms pay dividends, there is a great deal of variation in the firm-specific, day-specific dividend yield that can be used to identify the link between dividend yield and pre-tax return.

The difficulty of determining the marginal tax burden on capital-income flows is illustrated, in the dividend valuation context, by Miller and Scholes (1978) and several subsequent studies. Miller and Scholes (1978) noted that even for individual investors, who face current income taxation on dividends but not capital gains, it was possible for the effective tax burden on dividends to be low. They noted that IRS rules restrict tax-deductible individual borrowing to the larger of $10 000 or total capital income, which includes cash dividends. Receiving another dollar of cash dividends would therefore relax the borrowing constraint on an individual, and this could in effect make dividend income untaxed at the margin. Peterson, Peterson and Ang (1985) and Feenberg (1981) have explored the importance of this specialized tax provision, and they suggest that it does not play an important role in the dividend valuation of many households.

The interpretation of empirical evidence on the price movements of shares around their ex-dividend days depends critically on whether or not the shareholders who "typically" hold the firm are also holding the shares and setting prices around the ex-day. Allen and Michaely (1995) distinguish between "static clientele" theories in which long-term investors determine the ex-dividend day pricing relationships and "dynamic clientele" models in which investor clienteles in a given firm may be different on the ex-day and other days. Many studies of ex-day share-price movements implicitly assume that clienteles do not change over time. In this case the share-price decline around the ex-dividend day, when scaled by the dividend payment, may provide information on the marginal dividend and capital-gains tax rates on a firm's long-term investors. In more plausible models with time-varying clienteles, however, such inferences are more difficult.

Elton and Gruber's (1970) seminal study of ex-dividend-day price movements is an example of a study assuming static clienteles. This study found clear evidence that the share-price decline on ex-dividend days was smaller than the amount of the dividend payment, particularly for firms with relatively low dividend yields. They interpreted this finding as demonstrating that investors form clienteles on the basis of their tax rates, and that high-yield firms attract investor clienteles with low marginal tax rates on dividend income. Their analysis was premised on the view that ex-day

share-price movements reflect the tax rates of the long-term holders of the corporation's shares.

Elton and Gruber (1970) analyzed returns using monthly returns data, and other studies using monthly returns have reached different conclusions regarding the valuation of dividends and capital gains. Miller and Scholes (1982) argued that previous empirical findings such as those in Elton and Gruber (1970), which suggest that share prices decline by less than the value of dividend payouts, had been marred by statistical biases. They presented evidence using monthly stock returns over a long sample period that suggested that biases such as the coincidence of ex-dividend days and dividend announcement days, which would raise share prices, could account for a spurious positive relationship between pre-tax returns and dividend yields.

Gordon and Bradford (1980) also present evidence that is consistent with the Miller and Scholes (1982) conclusion in monthly data: they do not reject the null hypothesis that dividends and capital gains are valued equally. They study monthly stock returns over the period 1926–1978. One important innovation of their study is its explicit use of the after-tax capital asset pricing model to motivate the estimation strategy. This involves linking the after-tax asset-demand equations derived in the last section, with no limits on individual investor short sales, with information on asset supplies. The resulting equilibrium return relationship is

$$r_i + \nu * d_i - r_z = \beta_i * (r_m + \nu * d_m - r_z),$$

where r_i represents the pre-tax return on security i, d_i the dividend yield on security i, r_m and d_m the analogous return concepts for the market portfolio, and r_z the return on a zero-β portfolio.

The crucial parameter in this expression is ν (α in Gordon and Bradford's (1980) notation). It denotes a weighted average of the relative tax burdens on dividends and capital gains on different households:

$$\nu = \sum_h \frac{s_h}{\gamma_h} * \frac{1 - \tau_{div,h}}{1 - \tau_{cg,h}},$$

where s_h denotes the share of household h's wealth in total household wealth, γ_h denotes relative risk aversion for household h, and the tax parameters are defined as usual. This expression indicates that the relative price of dividends and capital gains is the same for all securities, and that it is determined as a weighted average of marginal tax rates with weights increasing in a household's wealth, and declining in its relative risk aversion. This expression implies that in the absence of short-selling constraints, the relative valuation of dividends and capital gains for all firms should be the same. This prediction is not consistent with a substantial body of empirical evidence that suggests the presence of dividend clienteles, and it raises questions about which of the various assumptions underlying this expression need to be relaxed.

Gordon and Bradford (1980) perform empirical tests of the model described above, under the maintained assumption that all firms face the same relative valuation of

dividends and capital gains. They find evidence of substantial fluctuation in the relative valuation of dividends and capital gains across five-year intervals, but they do not reject the null hypothesis that on average, this relative valuation is equal to unity. They also find evidence that the relative valuation of dividends and capital gains tends to move in tandem with Tobin's q, which is an empirical pattern that has not yet been investigated in subsequent research.

While empirical studies relying on monthly returns data find mixed evidence on the relationship between dividend yields and pre-tax returns, studies using daily returns tend to find clearer evidence that for many companies, dividends appear to be valued less than capital gains. One notable study using daily data is Barclay's (1987) analysis of ex-dividend-day pricing before the adoption of the federal income tax in 1913. His results suggest that share prices declined by approximately the full amount of their dividend payouts before 1913, while in the early 1960s, the comparison period he considers, prices declined by less than the full amount of the dividend. This finding is consistent with taxation playing a key role in determining ex-day pricing. Auerbach (1983) also presents evidence that v, as defined above, does not equal unity when it is estimated using daily data for the 1962–1977 period. There are some puzzles in these data, however. Eades, Hess and Kim (1984) study ex-dividend-day price movements in the United States, and they find that even for some distributions that are taxed in the same way as capital gains, the market seems to value the payouts less than dollar-for-dollar. This suggests that taxes may not be the only factor influencing returns around the ex-day. Michaely (1991) also presents evidence that is difficult to reconcile with the standard tax-based explanation of ex-dividend-day price movements. He finds no evidence that TRA86 affected the magnitude of price declines around ex-dividend days, even though this tax reform affected marginal tax rates for many investors.

Other higher-frequency comparisons of ex-day pricing around substantial tax reforms, and in other nations, reach varied conclusions regarding the effects of taxation and tax changes on ex-day pricing. Poterba and Summers (1984), for example, present evidence that the integration of the British corporate and personal income taxes was associated with a change in ex-dividend valuation. Morgan and Thomas (1998) and Bell and Jenkinson (2000) are more recent studies of ex-dividend pricing behavior in the United Kingdom, with different conclusions about the importance of tax considerations. Lakonishok and Vermaelen (1983) and Booth and Johnston (1984) study the effect of the 1971 Canadian tax reform that introduced capital-gains taxation. Green and Rydqvist (1999) present an intriguing analysis of ex-day pricing for lottery bonds in Sweden, and conclude that prices move as an after-tax ex-day model would suggest.

One interpretation of the rather mixed empirical findings in the ex-dividend-day pricing literature is that they are confounded by high-frequency fluctuations in shareholder clienteles, which implies that ex-day pricing does not reflect the stable, long-term clienteles in particular securities. A number of recent studies, beginning with Kalay (1982), have focused explicitly on the changes in ownership that take place around ex-dividend days. Kalay (1982) suggested that the bounds on short-term

profitable trading opportunities, which are a function of transaction costs, were the primary determinant of ex-dividend price movements. Several recent studies, including contributions such as Bali and Hite (1998), Bhardwaj and Brooks (1999) and Frank and Jagannathan (1998), have considered the extent to which market microstructure issues or other factors that are not related to taxes can explain the observed pattern of pricing.

The degree to which firms experience high-frequency changes in their tax clientele has also attracted substantial attention. Michaely and Vila (1996) document that volume around ex-dividend days is substantially greater than on average days, suggesting that some clientele changes are taking place. Karpoff and Walkling (1988, 1990) suggest that some investors engage in "dividend capture", trading in dividend-paying stocks around their ex-days. Koski (1996) also analyzes the short-term trading question, and devotes particular attention to the trading incentives facing corporations. She concludes that the combination of tax and regulatory changes in the early and mid-1980s sharply reduced the opportunities for profitable ex-dividend arbitrage that may have existed in the early 1980s. Koski and Scruggs (1998) focus on a single time period, 1990–1991, and present evidence of cross-sectional variation in the pattern of trading around ex-dividend days, with particular support for greater trading by securities dealers in stocks that have high dividend yields. Eades, Hess and Kim (1994) track the time-series fluctuation in the ex-day return patterns for high-yield securities, and they argue that these patterns are consistent with less dividend capture during periods when the costs of such trading was higher.

Research on high-frequency clientele changes suggests that there may be some incentive for such trading for securities with very high dividend yields. Recognizing the role of dividend-capture traders in determining share prices around ex-dividend days has been an important research insight of the last fifteen years. For many stocks with smaller dividends, however, ex-day arbitrage does not generate high returns because the transaction costs are too large to make such trading profitable. For these firms, the ex-day price movement may represent the relative valuation of dividends and capital gains by long-term investors. There may be more stable, longer-term clienteles in the ownership of these firms; Dhaliwal, Erickson and Trezevant (1999) present some empirical evidence on the importance of dividend policy in affecting the ownership of firms. For firms without short-run changes in dividend clienteles, the balance of evidence suggests that dividends are valued less than capital gains. This finding raises a puzzle with respect to corporate financial policy: the perennial question of "why do firms pay dividends"? Black (1976) concisely poses this puzzle, and Auerbach (2001) offers a review of contemporary work.

3. Taxation and asset sales

The preceding section considered the influence of taxation on the set of assets that individuals choose to hold in their portfolios, and the fraction of their wealth that they

choose to hold in different assets. The theory of portfolio selection underlying that discussion is well developed. This section considers a different set of issues: the effect of taxation on decisions about when to buy and sell assets. The tax which has the greatest influence on this decision is probably the capital-gains tax, and while there is a voluminous empirical literature directed at measuring the effect of capital-gains taxation on asset sales, the theoretical literature that underlies this work is poorly developed. There is no generally accepted model of why investors choose to sell assets, so it is difficult to embed the literature on capital-gains taxation in a theoretical framework in which welfare analysis is possible.

One of the reasons that capital-gains taxation has attracted so much research and policy attention is that it is one of the few situations in which there is a plausible empirical argument that reducing marginal tax rates may raise government revenues. Public-finance scholars have long recognized the possibility, popularized by economist and presidential advisor Arthur Laffer in the early 1980s as the "Laffer curve", that lowering marginal tax rates could increase total government revenue. There are few instances in which careful empirical research suggests that this possibility is a practical reality. Capital-gains taxation is one such case.

To understand the effects of the capital gains tax rate on current tax revenue, it is helpful to write revenue, R, as the product of the tax rate on realized gains, τ_{cg}, and the tax base, which equals realized gains (REALIZATIONS): $R = \tau_{cg} * \text{REALIZATIONS}$. The condition for a "Laffer effect", $dR/d\tau_{cg} < 0$, is

$$\frac{dR}{d\tau_{cg}} = \text{REALIZATIONS} + \tau_{cg} * \frac{d(\text{REALIZATIONS})}{d\tau_{cg}} < 0.$$

This can be rewritten in terms either of the elasticity of realized gains with respect to the marginal tax rate, $\eta_{\text{real},\tau} = d\ln(\text{REALIZATIONS})/d\ln(\tau_{cg})$, as $\eta_{\text{real},\tau} < -1$, or (in what may be a more natural elasticity to consider) the elasticity of realizations with respect to the after-tax income associated with a realization, $\eta_{\text{real},1-\tau} = d\ln(\text{REALIZATIONS})/d\ln(1-\tau_{cg})$, as $\eta_{\text{real},1-\tau} > (1-\tau_{cg})/\tau_{cg}$. Empirical work on the link between capital-gains realizations and marginal tax rates has focused on whether this condition is satisfied. It is important to recognize that this expression considers only the effect of the capital-gains tax rate on current capital-gain realizations. It is possible for changes in the tax rate at one date to affect realizations at other dates, and the associated revenue effects need to be considered in thinking about the effect of changes in capital-gains tax rate on the present discounted value of government revenues.

One of the difficult problems in evaluating the revenue effects of changes in the capital-gains tax rate is that even if realizations increase, it is possible that there are effects elsewhere in the tax system. If one source of realized capital gains is relabelling of other types of income, so that labor income declines when realized gains increase, then it is possible that a simple analysis of the link between capital-gains tax rates and capital-gains realizations may not fully describe the revenue effects of capital-gains tax reform.

3.1. Capital gains tax avoidance and loss-generation behavior

Before turning to the empirical evidence on capital-gain realizations and tax rates, it is helpful to describe optimal investor behavior in the presence of a realization-based capital-gains tax. Even a cursory review of optimal asset-trading strategies in an efficient capital market with such a tax may generate startling outcomes. In particular, astute investors and tax planners could in some cases generate negative capital-gains tax liability in all periods until their death, and then to use basis step-up to extinguish all of their lifetime capital-gains tax liability on accrued gains.

A number of studies have considered the optimal realization policy for an investor with a security that has an accrued gain or loss. Some features of optimal realization policy are straightforward. For example, an investor who holds securities that have declined in value since he purchased them can maximize the present discounted value of his tax deduction by realizing the loss immediately. In contrast, an investor with an accrued capital gain might defer the taxes on this gain by holding the gain for as long as possible, and ideally, until he dies and the value of the asset's basis is stepped up.

While there is agreement that depreciated assets should be sold immediately, there is disagreement concerning the appropriate treatment of appreciated assets. Constantinides (1984) argues that gains should be held until they must be realized to satisfy consumption needs or until basis step-up at death. Dammon and Spatt (1996), however, show that for low enough levels of transaction costs, it can be optimal to sell appreciated assets as well, as soon as their gains qualify for long-term status. The reason is that by selling the appreciated asset, and then repurchasing it, the investor can generate an opportunity for a short-term loss realization in the future.

A number of recent studies have developed new theoretical or empirical insights on the tax-timing issue. Leland (2000) examines the optimal trading rule in the presence of taxes, and he finds that with transactions costs there is a "region of inaction" in which investors do not trade, but that with low enough trading costs (or large enough gains and losses) the Dammon–Spatt strategies are still optimal. Dammon, Spatt and Zhang (2001) also explore optimal consumption behavior, and realization decisions, in the presence of capital-gains taxation.

Empirical research on the value of tax-timing options is limited. Brickley, Manaster and Schallheim (1991) investigate how the discount on closed-end mutual funds is related to the volatility of the underlying securities held by the fund. They argue that the value of the tax-timing option on the fund is worth less than the portfolio of tax-timing options on the individual securities, and that this differential should become larger as volatility rises. Their empirical results support this implication of the tax-timing analysis, although they might be consistent with other explanations as well. Chay, Choi and Pontiff (2000) also test the value of tax-timing options, in this case by studying the market value of forced capital-gain realizations. They conclude that the effective tax rate is about ninety percent of the statutory tax rate on realized gains. There has also been some work on tax-timing behavior in bond markets. Prisman,

Roberts and Tian (1996), for example, find evidence that investors take advantage of "tax-timing options" in the Canadian bond market.

One can move beyond the analysis of optimal realization policy for assets that an investor already holds to ask a more general question: can investors pursue portfolio strategies that will reduce their capital gains or income tax liability, and if so, what will these strategies involve? Several studies have noted that if investors can take large positions in securities with negatively correlated returns, they can *generate* capital losses that can be used to offset other types of capital income.

The simplest illustration of a loss-generation strategy, which has been explained in Constantinides and Scholes (1980) and Stiglitz (1983), involves two securities with perfectly negatively correlated returns. A long and a short position in the same security are a good illustration. At the beginning of a tax year, an individual purchases 100 shares of stock in Company X, and at the same time, he sells short 100 shares of this stock. This pair of transactions requires no commitment of net worth, at least in a frictionless financial market. The investor holds the two positions until the end of the tax year, at which time he either sells his shares or closes out his short position. Which transaction he undertakes depends on the performance of Company X's stock price over the year. If the stock has appreciated, the investor will have a gain on his long position in the stock, so he will close out his short position and generate a capital loss. If the stock price has fallen, however, he will sell his shares, thereby realizing a loss, and he maintains the short position. This strategy yields a certain capital loss in the current tax year, and a certain carry-forward of an accrued gain to the next year.

The transaction sketched above would not result in an allowable capital loss under current tax rules, because an investor with both a long and a short position in the same security would not be "at risk" in the underlying security. However, it is possible to pursue strategies similar to that described above using either two highly but imperfectly correlated securities, such as stocks in two oil or steel companies, or by using derivative securities. The attractiveness of strategies of this type depends critically on the transactions costs associated with establishing the various positions.

The degree to which investors pursue capital loss generation strategies is an empirical issue. Poterba (1987) presented data based on the 1985 IRS Sales of Capital Assets data file. These data show that less than one fifth of investors, and possibly only one tenth, realize the maximum deductible level of net capital losses, as the foregoing analysis would suggest. It is important to note that capital losses of more than $3000 cannot be deducted from ordinary income. Seyhun and Skinner (1994) find evidence consistent with Poterba (1987), in that relatively few investors appear to have net realized losses as we would expect if investors were using Constantinides–Scholes–Stiglitz strategies to generate losses. Auerbach, Burman and Siegel (2000) find that data for tax years in the 1990s suggest a much higher fraction of investors (as many as one third) reporting net capital losses. This may reflect a growth in investor sophistication, or a shift in the underlying distribution of returns on the assets that are being sold for tax purposes. Further work is needed to explain this shifting pattern.

One general difficulty with the literature on taxation and optimal trading behavior remains something of a mystery. Odean (1998) and Shefrin and Statman (1985) suggest that individual investors are reluctant to realize their losses, partly because there are psychological costs to acknowledging that one has participated in a loss-generating trade. Grinblatt and Keloharju (2000, 2001) further explore the factors that induce trading with a rich data set on Finnish investors; their data provide some support for the role of tax-loss trading by investors. Future work is needed to link this literature with the studies of what optimal realization behavior would be in perfect capital markets.

3.2. Asset turnover and the capital gains tax: empirical evidence

The empirical study that launched the modern literature on how capital-gains taxation affects realization behavior is Feldstein, Slemrod and Yitzhaki (hereafter FSY) (1980). This study used data from individual income-tax returns for tax year 1973 that were released as part of an IRS Statistics of Income – Sales of Capital Assets file. The key regression equation related a taxpayer's long-term capital gains on sales of corporate stock (LTG), divided by the taxpayer's dividend income (DIV, as a proxy for total holdings of corporate stock), to the taxpayer's marginal tax rate on realized capital gains. The estimation sample was limited to taxpayers who reported at least \$3000 of taxable dividends. The results are (with standard errors shown in parentheses):

$$\underset{}{\mathrm{LTG/DIV}} = \underset{(1.3)}{35.0} - \underset{(3.8)}{49.7} * \tau_{cg} + \underset{(0.35)}{0.18} * \mathrm{AGE65+} - \underset{(0.12)}{1.23} * \ln(\mathrm{DIV}) - \underset{(0.12)}{0.50} * \ln(\mathrm{AGI}).$$

These results imply that a ten-percentage-point reduction in the marginal tax rate on capital gains would raise the ratio of long-term gains, which averages 3.50 in the sample, by nearly 5. The estimates can also be interpreted in elasticity terms. Since the average value of the marginal tax rate on capital gains is 0.264, the implied elasticity is –3.75, so gains respond to tax rates by more than enough to generate revenue gains from reductions in marginal rates.

One of the critical empirical issues in studies of capital-gains realizations, as well as related taxpayer behaviors such as charitable giving or borrowing, is that the marginal tax rate on the last dollar of realized gains may be affected by the level of realizations. This induces a fundamental endogeneity between the independent variable of interest, the marginal tax rate on realizations, and the dependent variable. FSY (1980) tackle this problem by also constructing a *first-dollar* marginal tax rate on realized gains. This is a measure of the marginal tax burden assuming that the taxpayer had not realized any gains, and it is therefore independent of actual realizations. This marginal tax rate can either be used as the independent variable for the regression model above, or, as more recent studies have done, it can be used as an instrumental variable for the actual, last-dollar marginal tax rate. FSY report that including the first-dollar tax variable in their specification, in place of the last-dollar marginal tax rate, results in a coefficient

estimate of −37.1 rather than −49.7. This still implies a large elasticity of gains with respect to the tax rate.

The FSY study suggested that if marginal tax rates on realized gains were reduced from their levels in the early 1970s, the total revenue collected from the capital-gains tax would increase. This conclusion has been questioned, however, by a number of subsequent studies that have focused on some the empirical difficulties in estimating a realization elasticity from tax-return data.

One critical empirical difficulty that arises in any study of how current realized gains depend on the current marginal tax rate involves distinguishing transitory and permanent effects on realization decisions. There are several dimensions of this problem. One is that if a given household experiences year-to-year fluctuations in income, which are associated with fluctuations in marginal tax rates, the household may try to time capital-gain realizations to coincide with years of low marginal tax rate. This possibility was recognized by FSY (1980), but it was not possible to address this difficulty using only a single cross-section data set. If households engage in this type of retiming behavior, however, then the estimated elasticity of realizations with respect to marginal tax rates in a single cross-section may not indicate how a permanent reduction in the capital-gains tax would affect realization behavior.

A second dimension of the transitory–permanent problem arises when capital-gains tax rates are known to be changing in the near future. There may be substantial re-timing of realizations, and the short-run elasticity of realizations with respect to the marginal tax rate may be high, even if the long-run elasticity is low. The circumstances surrounding the Tax Reform Act of 1986 illustrate the potential importance of re-timing behavior. In that case, it was clear by mid-1986 that the top marginal tax rate on gains realized after January 1, 1987 would be 28 percent, while the top rate on gains realized before that date was 20 percent. The time series of realized long-term gains for the mid-1980s indicates the impact of such an anticipated capital-gains tax increase. These realizations, measured in $1986 billion for the five years beginning in 1983, were $129.8, $145, $171.2, $324.8, and $144.4. The empirical challenge posed by findings such as this is separating the transitory and permanent effects of capital-gains tax changes.

A number of studies have extended the FSY (1980) methodology by allowing for both permanent and transitory realization elasticities. Burman (1999), Gravelle (1994) and Mariger (1995) discuss a number of these studies. Auten and Clotfelter (1982) use a panel of tax returns for the period 1969–1973, and their empirical strategy involves the inclusion of both the current marginal tax rate on long-term gains, as well as the average of the individual's tax rates over the years in the panel data set. The statistical results suggest that there are important differences between the impact of the current tax rate, and the impact of the average or permanent tax rate, on realized gains. The estimated elasticity of long-term gain realizations with respect to the permanent tax rate is −0.37, while the estimated elasticity with respect to transitory fluctuations in the tax rate is −1.05. These findings suggest that the long-run realization effect of cutting the capital-gains tax rate may be smaller than that required

to increase revenues. Auten and Joulfaian (1999) present more recent evidence using panel data, and they also find substantial differences between the impact of permanent and transitory changes in tax rates.

One of the most widely discussed studies on capital-gains taxation and realization behavior is by Burman and Randolph (1994). They use a panel of tax returns for the period 1979–1983, and they use variation in marginal tax rates due to the state a taxpayer lives in, and thus the state income tax rate on reported gains, as a source of "permanent" tax-rate variation. Their basic empirical specification is given by

$$\mathrm{LTG}_t^* = X_t\alpha_0 + \tau_{\mathrm{cg,\,perm}} * \alpha_1 + \tau_{\mathrm{cg},\,t} * \alpha_2 + \tau_{\mathrm{cg},\,t-1} * \alpha_3 + \varepsilon_t.$$

LTG* denotes the desired level of long-term gain realizations; it can be negative, and the estimation relies on a Tobit estimator to handle truncation at zero. This specification allows for a separate effect of the permanent tax rate on capital gains ($\tau_{\mathrm{cg,\,perm}}$), the current tax rate on capital gains ($\tau_{\mathrm{cg},\,t}$, which one can alternatively view as the deviation of the current tax rate from the permanent level), and the lagged deviation of the tax rate from its permanent level ($\tau_{\mathrm{cg},\,t-1}$).

The empirical findings suggest a large transitory elasticity of capital-gains realizations with respect to the marginal tax rate, with an elasticity estimate of –6.42 (0.34) in the base case. The estimate of the realization elasticity with respect to permanent tax changes, however, is much smaller, and it is statistically insignificantly different from zero: –0.18 (0.48). Thus these findings confirm the earlier suggestion that the long-run realization elasticity may fall short of the value needed to imply that reducing capital-gains tax rates would raise revenue.

The debate on the effect of capital-gains taxation on gain realizations is likely to continue, since one can raise objections to essentially all of the existing empirical work. For example, Burman and Randolph's (1994) identification using cross-state differences leads to questions about whether state of residence is itself endogenous. There is some empirical evidence that wealthy, elderly taxpayers are somewhat sensitive to capital income and estate tax rates in choosing their state of residence; this makes it difficult to evaluate the Burman–Randolph results. Moreover, for addressing the question of how taxpayers would respond to an actual change in the federal tax rate on long-term gains, it is important to ask what taxpayers would believe about the likely permanence of such a change. If taxpayers viewed such a change as transitory, then the short-term realization effects could be as indicated by the transitory, rather than permanent, capital-gains tax rate variables in the foregoing specification.

In addition to the cross-sectional and panel-data studies described above, there have been some studies of aggregate capital-gains realizations and the effect of marginal tax rates using time-series data. The substantial literature on this issue is surveyed by the US Congressional Budget Office (1988) report on capital gains taxation. Auerbach (1988) represents the most careful analysis of the time-series record to date. The findings in this literature parallel those in the studies that have used taxpayer data:

they show clear evidence of high-frequency effects of the capital-gains tax rate on the flow of realizations, but much weaker evidence that permanent changes in capital-gains tax rates affect the flow of realizations. Auerbach (1988), for example, finds that when only the contemporaneous capital-gains tax rate is included in a regression equation for capital-gains realizations estimated over the period 1955–1985, the coefficient on the tax rate variable is −4.3, with a *t*-statistic of −2.4. When the current tax rate and the tax-rate change from the previous period are both included in the specification, however, the coefficient on the contemporaneous tax rate falls to −1.8 (*t*-statistic −0.7), and the coefficient on the tax-rate change variable is −1.8 (*t*-statistic −0.9). These findings illustrate the limited amount of information in the time-series evidence, and the sensitivity of time-series findings with respect to minor changes in specification. Eichner and Sinai (2000) show that even with a longer time series running through the mid-1990s, the elasticity estimates are still very sensitive to particular sample periods, especially the inclusion of the years 1985–1987.

The lack of robust results from the time-series analysis is unfortunate, because in some ways the aggregate data may be the best source of information on the effect of tax changes on realizations. It describes the effect of tax changes when all of the general equilibrium effects of the tax cut, such as asset-price changes and changes in the advice of financial intermediaries, are allowed to take place. Cross-sectional evidence does not provide any information on the potential magnitude of effects through these channels.

Before leaving the discussion of how capital-gains taxation affects the flow of gain realizations, several additional points deserve comment. First, there are strong reasons to think that the effect of a capital-gains tax change on realizations will depend on the past history of asset returns and tax rates. Cutting capital-gains tax rates after they have been high, and after assets have risen sharply in value, is likely to have a larger effect on the flow of realizations than a similar-sized reduction in rates starting from a lower base tax rate or after a less robust period of asset returns. There is an emerging literature, illustrated for example by Shackelford and Verrecchia's (1999) analysis of capital-gains tax rates, on how taxpayers respond to anticipated changes in taxes. One interesting finding, reported in Auerbach and Siegel (2000), is that long-run responses to changes in the tax code, as well as short-run "timing" responses, may vary across taxpayers with different levels of tax sophistication.

Second, virtually none of the previous research on the capital gains tax has considered how changes in this tax might affect the reporting of non-capital-gains income. In particular, there is little work on whether there is substantial re-labeling of ordinary income as capital-gains income when the capital-gains rate is below the rate on interest, dividends, and wages. One of the primary activities of tax planners is transforming ordinary income into capital gains; the key issue is how important this is at an aggregate level. Third, there has been relatively little research on the degree to which realization elasticities vary across asset categories. The mix of assets generating gains has shifted over time, with the fraction of gains due to sales of corporate stock rising in the last decade. For less liquid assets, such as commercial real estate, effect

of changing the capital-gains tax rate on realizations may be smaller than for more liquid assets such as corporate stock.

Finally, the foregoing discussion has not discussed in any detail one of the most important features of the US capital-gains tax, which is the "basis step-up at death" provision. A taxpayer who dies with an appreciated asset can leave this asset to an heir, and the heir will inherit the asset with a new, "stepped-up" basis equal to the asset's value at the time of the first person's death. Basis step-up effectively extinguishes the tax liability on capital gains that accrued during the decedent's lifetime. This tax provision has two important effects. First, it reduces the effective capital-gains tax rate to a rate substantially below the statutory rate; Bailey (1969) estimates that this provision reduced the effective tax burden on capital gains by about 50%. Protopapadakis (1983) presents related calculations using the rate of capital-gain realizations to estimate the effective tax rate. Second, for elderly individuals with relatively short life expectancies, the basis-step-up provision creates a transitory and predictable fluctuation in the capital-gains tax rate, and it may lead to particularly pronounced "lock-in" effects for those near the end of the lifecycle.

There is little empirical evidence, however, on the effect of basis step-up on asset sales; this is an issue that deserves further analysis. Poterba and Weisbenner (2001a) present evidence on the distributional effects of shifting from the current estate tax, with basis step-up, to a system that included unrealized capital gains in the taxable income of decedents for their final year. There is some experience in Canada in the early 1970s with a shift from an estate tax to a capital-gains tax at death, described in Bossons (1972), but this has not yielded insights on the behavioral effects of such a change.

3.3. Taxation and the January effect

One issue involving portfolio behavior and taxation, which has attracted some attention in both financial economics and public finance, is the link between tax-motivated investor trading and the so-called "January effect" in stock returns. The "January effect" is the systematic finding that the average return on common stocks is higher in January than in any other month, as least in the US equity market. This effect is somewhat more pronounced among small stocks and stocks that have experienced losses in the previous year. While there is suggestive evidence that investors sell shares with losses as the year-end approaches, as efficient tax management would dictate, there is only limited empirical evidence linking this trading to the January effect, or showing that it is large enough to explain the abnormal January returns.

Badrinath and Lewellen (1991) is the clearest study of year-end tax-motivated trading. This study uses transactions data from individual accounts at a major brokerage firm. It finds a higher concentration of transactions that generate losses in December than in any other month. This evidence is consistent with the studies that consider aggregate volume in individual companies, such as Dyl (1977) and Slemrod (1982), and relate it to the firm's recent return performance. While there is usually a

negative relationship between trading volume in January and the security's historical return, Bolster, Lindsey and Mitrusi (1989) suggest that this pattern reversed in 1986, a year when rising capital-gains tax rates made it attractive to realize gains before year-end. This evidence all points to an important link between tax considerations and investor trading decisions. Seida and Wempe (2000) move beyond an analysis of volume by using intra-day transaction data that makes it possible to identify stock sales by individual investors. Their findings also suggest a sharp increase in sales of appreciated assets in late 1986.

Three recent empirical studies provide further evidence linking tax considerations with end-of-year stock trading and stock returns. Sims (1995) shows that firms that have experienced losses during the calendar year that is about to end experience more negative returns just before the end of the year than other stocks do. This is consistent with a "return rebound" for the shares in these firms after the turn of the year. Poterba and Weisbenner (2001b) show that the relationship between past stock returns and January returns is a function of the precise features of the capital-gains tax. In particular, changes in the definition of short-term and long-term losses appear to affect the link between past returns and January returns. Grinblatt and Moskowitz (2000) present additional evidence that confirms this general finding. The discovery that parameters of the tax code affect the relationship between lagged returns and current returns provides some support for the role of tax-loss trading in generating abnormal January returns. The leading alternative hypothesis to explain this pattern, "window dressing" on the part of institutional money managers, would not suggest such a pattern.

One interesting extension of the "January effect" literature is the possibility of a "November effect" associated with trading at the end of the tax year for mutual funds. Bhabra, Dhillon and Ramirez (1999) suggest that as mutual funds have become more important investors in the equity market, there has been a growing pattern of return abnormalities around the end of their tax year.

3.4. The welfare effects of capital-gains taxation

While there is a large empirical literature directed at measuring the effects of capital-gains taxation on investor behavior, there is relatively little theoretical work addressed to the welfare effects of realization-based capital-gains taxation. The capital-gains tax contributes to the overall tax burden on capital income, and the general analysis of the welfare cost of capital-income taxation in the spirit of Feldstein (1978) and Atkinson and Sandmo (1980) is therefore relevant. In addition, however, the unique behavioral effect of a realization-based capital-gains tax is that it creates disincentives to sell appreciated assets, and it thereby creates a "lock-in" effect. The welfare consequences of such lock-in have only begun to be studied.

There have been several attempts to develop models of how capital-gains taxes affect asset realization decisions. Balcer and Judd (1987), for example, explore the optimal structure of asset purchase and liquidation in a lifecycle model. They abstract from

uncertainty about rates of return on different assets, and assume a constant rate of asset-price appreciation in all periods. In this setting, they show that it is optimal for an investor to liquidate assets with the highest basis (purchase price) at any point in time; these will be the most recently purchased assets. In addition, they show that it is impossible to refer to "the" effective capital-gains tax rate, because the burden of a realization-based tax depends critically on holding period and the pre-tax appreciation of the underlying asset. Balcer and Judd (1987) do not present any explicit calculations of the welfare cost of capital-gains taxation, and their model is not well suited to studying the problem of lock-in across securities with different historical returns.

Kiefer (1990) represents a second attempt to study capital-gains taxation and its effect on investor behavior. This paper uses a simple simulation model, in which investors share expectations about the prospective rate of return on assets that they do not own, but have heterogeneous expectations about the rates of return on assets they do own. This structure determines which investors will sell assets at a given point in time, and it can be used to study the hypothetical reaction of investors to a change in the capital-gains tax rate. Unfortunately, the link between this simple model and actual investor behavior is unclear, and the simplified structure of the model makes it difficult to calibrate it. There is also no attempt to address the welfare consequences of realization-based capital-gains taxation.

Auerbach (1992) also explores the welfare cost of capital-gains taxation in a stylized three-period model. The model suggests that the equivalent variation associated with a shift from the current realization tax system to an equal-revenue accrual-based system, could be equal to several percent of household wealth. The analysis also indicates that by reducing the lock-in effect, a switch to accrual taxation could depress personal saving in the years surrounding the tax transition.

Kovenock and Rothschild (1987) present another analysis of portfolio lock-in and its welfare consequences. They consider an investor's expected utility from following different portfolio investment strategies in a multi-period investment problem. One strategy, the optimal strategy in a world without realization-based taxes, is to rebalance the portfolio weights in every period to reflect current information on prospective returns. The other strategy, which may prove optimal with high rates of realization-based taxation, is to follow a "buy and hold" strategy without any rebalancing. The paper does not derive an optimal portfolio adjustment strategy in the presence of realization-based taxation, but it does consider the types of strategies that would be more attractive with realization-based taxes than without them. Kovenock and Rothschild (1987) show that investors experience lower expected utility when they do not rebalance their portfolios. As in Balcer and Judd (1987), the focus in presenting results is on the comparison of effective tax rates rather than on more direct welfare comparisons, but the results are suggestive about the costs of realization-based taxation. One limitation of the analysis is that it does not endogenize the decision of whether or not to sell a given asset. If the expected utility gains from realizing an appreciated asset, paying capital-gains tax, and re-investing in a

balanced portfolio are positive, investors should do this, yet the paper does not allow this option.

There is very little empirical evidence on the extent to which investors are locked-in to particular assets. One notable study by Landsman and Shackelford (1995) investigates how an investor's basis in the stock of a single firm, R.J. Reynolds, relates to the price at which they tendered the stock to a takeover bidder. This study finds that investors who had purchased RJR stock at low prices were more likely to wait until later in the takeover process before selling out; this supports the view that capital-gains basis can affect the reservation price that individual's demand for selling their shares. Reese (1998) presents a related, and clever, test of how capital gains affects trading behavior. He studies recent initial public offerings (IPOs) of common stock, so that he knows the maximum possible holding period for an investor in the security. He finds that for IPOs that appreciate in their first year of trading, there is a substantial increase in trading just after the IPO has been traded for one year. There is an analogous effect just *before* the IPO reaches the one-year mark for shares that have declined in value. This pattern is consistent with investors holding shares with accrued gains longer than they might otherwise to qualify those shares for a long-term gain. Klein (1999) more generally explores, in a theoretical setting, the link between locked-in investors and the required return on different securities. He shows that when a substantial number of investors are locked in to an asset, the expected return on that asset may be lower than the expected return on other securities.

Another related study is Burman, Wallace and Weiner's (1997) analysis of sales decisions by homeowners. This paper presents weak evidence that in the United States, the probability of homeowners with accrued capital gains selling their homes and purchasing smaller homes rises after they reach age 55. During the period of their data, those who sold homes with gains before age 55, and who did not roll the gains over into a new home, had to pay capital gains tax on the full amount of the gains. After age 55, $125 000 of capital gain could be excluded from taxation. This represents another example of lock-in behavior, but its welfare effects have not been explored. The tax rules that generated this lock-in effect were modified in the Taxpayer Relief Act of 1997. Housing capital gains of less than $500 000 are no longer subject to capital-gains tax. This change has presumably reduced the potential role of lock-in in the residential real estate market.

A small but expanding body of research has documented an effect of capital-gains taxes on asset prices, particularly the prices of common stock. Amoako-Adu, Rashid and Stebbins (1992) find that the introduction of the Canadian capital-gains tax was associated with substantial asset revaluations. Lang and Shackelford (2000) and Shackelford (2000b) present evidence for the United States, showing that the stocks that were best positioned to benefit from lower capital-gains tax rates rose the most when legislators moved toward capital-gains tax reduction in 1997. Parallel evidence on the capitalization of the dividend tax burden is shown in Ayers, Cloyd and Robinson (2000), for the 1993 tax change in the United States, and in Poterba and Summers (1985), for dividend tax changes in the United Kingdom. Blouin, Raedy

and Shackelford (2000) investigate the incidence of the capital-gains tax burden on existing shareholders in a company that experiences an exogenous shift in demand; they find that new buyers must compensate existing holders, at least in part, for their tax burden.

3.5. The securities-transactions tax and capital market equilibrium

Taxes on realized capital gains are the tax policy instrument that are most often discussed in studies of asset turnover, but they are not the only tax that can affect the decision to sell assets. Another tax instrument that periodically attracts substantial policy discussion is the securities-transactions tax (STT) which Tobin (1978) proposed as a device for throwing "sand in the gears" of the markets in which financial securities are traded. Tobin's basis for suggesting such a tax was that some speculative trading imposes negative externalities on the financial system, so that a STT could be viewed as a corrective Pigouvian tax. Recent research on the role of "noise traders" in securities markets has provided a theoretical framework for considering the potential externalities associated with trading behavior, and in this context, Summers and Summers (1989) suggest that there might be welfare gains from adopting a securities-transactions tax. The substantial volume of financial transactions on the major stock markets, and in markets for derivative securities, has drawn policy makers to the STT. Assuming, as is very unlikely, that the volume and location of trade was not affected by a transactions tax, the revenue potential of the STT is substantial.

Most of the debate on the welfare gains or losses from adopting a securities-transactions tax involves a comparison of alternative theoretical models. Schwert and Seguin (1995) provide a valuable introduction to this research. Because relatively few nations have imposed, or changed, securities-transactions taxes in recent history, there is little empirical evidence on the effect of such taxes. Sweden provides a notable exception to this lack of tax variation: in 1984, Sweden imposed a 50 basis point tax on all purchases and sales of equities, and in 1986, the one-way tax rate was raised to 100 basis points. Umlauf (1993) provides a careful analysis of the impact of the Swedish STT. He shows that when Sweden raised its securities-transfer tax, trading volume in Swedish securities *in Sweden* declined, but that much of this volume moved offshore, where trades could be consummated without paying the transactions tax. Lybeck (1991) estimates that the elasticity of trading in Swedish money-market instruments within Sweden, with respect to the transactions tax rate, is approximately minus three. Campbell and Froot (1995) report that the revenue collected by the Swedish STT was less than one twentieth of the initial revenue projections.

Hubbard (1995) discusses more generally the extent to which securities trading is likely to move "offshore", or to move to different types of securities, as a result of a unilateral national tax on securities transactions. Because the location of securities transactions is a relatively elastic decision variable, changes in transaction taxes are likely to have substantial effects in altering the location of trade. Thus it is possible

to imagine securities-transaction taxes that reduce the domestic volume of trade but have no effect on the total international volume of trade in a given security.

4. Taxation and the markets for particular financial products

Many of the issues and research questions involving taxation and portfolio structure are specific, involving particular financial institutions, assets, or financial products. This section considers a number of these issues, with an emphasis on topics that are likely to attract growing attention in the future.

4.1. The tax-exempt bond market

One of the most direct applications of the theories of taxation and portfolio choice described above is with respect to the market for tax-exempt securities. In the United States, most of the bonds issued by state and local governments are exempt from federal interest-income taxation. If the risk characteristics of these bonds were identical to those of taxable bonds, for example Treasury securities, then simple models of portfolio equilibrium would suggest that investors in high-tax brackets would hold these securities. The lowest-tax-bracket individual holding tax-exempt bonds would be the "marginal investor" in these bonds, and his marginal tax rate would determine the yield spread between taxable and tax-exempt interest rates: $R_{\text{exempt}} = (1 - \tau_{\text{marginal}}) * R_{\text{taxable}}$. Auerbach and King (1983) and McDonald (1983) discuss this prediction in the context of clientele portfolio models like those presented above.

The observed yield spread between taxable and tax-exempt bonds in the United States, particularly at long maturities, has often been much smaller than this analysis would suggest. Kochin and Parks (1988) suggest that there have been periods when the long-term yield spread ($R_{\text{exempt}} - R_{\text{taxable}}$) has been so narrow that implied future short-term rates on tax-exempt bonds have been higher than comparable short-term interest rates on taxable bonds. This is not to suggest that taxation does not affect the yield spread on taxable and tax-exempt bonds. The event-study evidence, provided for example by Poterba (1986b) and Slemrod and Greimel (1999), demonstrates that tax reforms do affect the yield spread between taxable and tax-exempt bonds.

Various explanations for observed yield differentials have been suggested, but none have completely explained the observed pattern. Fortune (1988) discusses this work in some detail. Some studies have suggested that risk differences may explain narrow yield spreads, but Chalmers (1998) presents data on tax-exempt bonds that are effectively riskless, because their future payouts have already been funded by the borrower. He concludes that risk adjustments cannot explain the relatively narrow yield spread between taxable and tax-exempt securities.

Green (1993) emphasizes that fully-taxable investors would not compare the tax-exempt bond rate with that on taxable bonds that yield only interest income, but rather

would construct a taxable-bond portfolio of bonds that sell below their par values and therefore generate some capital gains as well as some interest. This suggests that the implicit interest-income tax rate on long-term bonds is higher than the foregoing calculation would suggest. The tax rates of the investors who hold fully taxable bonds are lower than those of investors who hold other (less heavily taxed) bonds, and who are comparing such bonds with tax-exempt bonds. This calls into question the standard "implicit tax rate" that is also computed based on the yields on fully taxable and tax-exempt par bonds. The observation that investors may form tax-based clienteles in the bond market does not apply only to tax-exempt bond markets. Green and Odegaard (1997) present evidence of clientele formation in the market for US Treasury bonds.

Evidence on the ownership of tax-exempt bonds is broadly consistent with tax-based clientele models, although there are some puzzles. Poterba and Samwick (2002) show that household tax rates are strongly correlated with the likelihood that the household owns tax-exempt bonds and with the portfolio share in such bonds. Feenberg and Poterba (1991) present information from 1988 individual income tax returns, on which individuals were asked to report their tax-exempt interest income even though this income was not included in the federal income tax base. The results illustrate that households in the lowest federal marginal income tax bracket received roughly one fifth of the tax-exempt interest that was received by households in 1988. Similar tabulations for more recent years confirm this finding. Why such individuals hold tax-exempt bonds is an open question. It might be because these are illiquid securities that they never chose to purchase, but instead received as an inheritance. It might be that their marginal tax rates fluctuate from year to year, and that when they are observed in a cross-section, their tax rates are transitorily low. This is an empirical issue that can be resolved with further study.

4.2. Taxation and mutual funds

One of the most significant changes in the structure of household portfolios in the United States during the last two decades has been the decline in direct individual ownership of corporate stock, and the corresponding rise in stock ownership through intermediaries such as mutual funds. The Investment Company Institute (1999) reports that 41 percent of US households own mutual funds, either through a retirement plan or through a directly taxable account. The rapid expansion of mutual-fund ownership during the 1990s has been one of the important forces behind the growth of stock ownership.

The growth of mutual funds is something of a puzzle from the standpoint of both tax-efficient investing and pre-tax return management. From a tax perspective, investors who hold assets through a mutual fund forego the opportunity to manage their capital-gains realizations. They also forego the opportunity to select assets with a mix of dividends and capital-gains income that best suits their tax status. From the standpoint of pre-tax returns, Gruber (1996) explains that the puzzle associated with mutual funds is that their average return is substantially below that of most stock-market

indices, largely as a result of transaction costs and expenses. While mutual funds do offer individuals a convenient and time-efficient way to manage their assets, and they perform a set of record-keeping functions that may also be valuable to investors, it remains unclear whether these advantages justify the tax and expected return penalty often associated with these investments.

Research on taxation and mutual-fund investments has focused on two issues. The first concerns the measurement of after-tax returns on mutual funds, and the extent to which mutual-fund investors consider after-tax returns in allocating their assets. The second concerns the behavior of mutual-fund managers, particularly with respect to capital-gain realization decisions.

With respect to the measurement of after-tax returns, Dickson and Shoven (1995) show that the focus on pre-tax returns can yield a misleading measure of how a mutual fund ranks relative to other comparable funds, and they recompute performance on an after-tax basis. Jeffrey and Arnott (1993) and Arnott, Berkin and Ye (2000) present evidence on the substantial tax cost of holding many actively managed equity mutual funds. Bergstresser and Poterba (2002) build on this work by studying the link between after-tax returns and the inflows of funds to mutual funds. They find that both the pre-tax return and the tax burden on a fund are related to the inflow, the former with a positive and the latter with a negative effect.

The taxation of mutual-fund returns is complicated, at least in the United States, by a set of rules that were specified in the Investment Company Act of 1940. If an individual purchases an individual stock and the stock rises in value, the individual is not liable for capital-gains tax until he sells the stock and realizes the gain. With a mutual fund, however, the key realization decision is that of the fund manager, not the individual investor. When a fund sells assets and realizes a capital gain, this gain is immediately passed-through to investors holding shares in the fund. Thus even if the investor does not sell his shares in the mutual fund during the year, he could be liable for capital-gains taxes. Funds differ substantially in the degree to which they realize gains, and therefore in the size of the potential tax burden that they impose on long-term investors in the fund.

The pass-through rules for mutual-fund capital-gains also raise the possibility that an investor can purchase a fund, experience no price appreciation on the shares in the fund during a given tax period, but still face capital-gains tax liability as a result of the fund investment. Many funds have an "overhang" of unrealized capital gains. This overhang is the result of unrealized gains in past years. Whenever the fund manager decides to sell assets with unrealized gains, these gains will be distributed on a *pro rata* basis to all shareholders in the fund. Someone who has just purchased the fund could therefore face a capital-gains tax bill even though this investor might not have earned any capital gains since buying the fund. This capital-gains tax liability alters the timing of taxes relative to what they would be if the investor's behavior, rather than the manager's, determined the realization date for gains. The new investor's tax basis in the mutual fund will be increased by any distributed gains on which he pays taxes. When

he does sell his shares, he will therefore be liable for a smaller capital-gains tax bill than he would if his own realization decisions were the sole determinant of his taxes.

Managers deciding to sell one asset and buy another can trigger gains in a mutual fund, but realizations can also be generated by redemption decisions on the part of some fund shareholders. Within a mutual fund, redemptions by one set of investors impose externalities on the other investors. Dickson, Shoven and Sialm (2000) explore a number of strategies that mutual funds might use to reduce the externalities that investors impose on each other as a result of their redemption decisions. These include exit charges, which can be used to compensate the shareholders who must bear the increased tax burden, or the creation of "tiered" mutual funds that would avoid co-mingling funds that were invested in the mutual fund at different dates.

When fund managers decide to realize gains, they deprive their investors of the benefits of deferring capital-gains taxes into the future. This raises the second major question about taxation and mutual-fund behavior: to what extent do fund managers consider their taxable investors' taxes in managing their assets? Dickson and Shoven (1994) show that by following simple realization strategies, such as always selling the high-basis stock in any security that they wish to reduce their holdings of, managers could significantly increase their after-tax returns. They would also increase the unrealized capital-gain "overhang" in their funds. A central issue is therefore whether managers try to avoid building up capital-gains overhang, or whether they accumulate unrealized gains to reduce their investors' current tax burden.

Several studies have addressed this issue. Huddart and Narayanan (2000) report some evidence of tax-sensitive trading on the part of mutual-fund managers. They find that there are differences in the year-end realization behavior of institutional money managers at untaxed institutions and at mutual funds, and that mutual-fund managers do appear to consider, at least to some degree, the tax burden that realizations will impose on their shareholders. Barclay, Pearson and Weisbach (1998) argue that fund managers have an incentive to realize gains and avoid a large overhang, even if this is not the way to maximize after-tax returns for existing fund shareholders, because this maximizes the fund's appeal to prospective investors. They argue that because mutual-fund managers are usually compensated based on their assets under management and their pre-tax return performance, they are concerned more with attracting new money into their fund than with maximizing the after-tax return to existing investors. This analysis does not consider the dynamic consistency problems associated with following a strategy that is attractive to new investors at the expense of old investors, i.e. the fact that new investors will be old investors in the future. Kraft and Weiss (1998) present intriguing evidence on the difference in realization behavior between open-end and closed-end mutual-fund managers. They find that closed-end fund managers, who do not need to consider the attractiveness of their shares for prospective investors, time their tax realizations in a fashion that minimizes tax burdens for individual investors, while most open-end fund managers do not.

The growing interest in the after-tax return to mutual fund investments has led to some changes in the mutual-fund marketplace. Khorana and Servaes (1999) find

evidence that when the existing mutual funds in a market niche are characterized by high levels of unrealized capital gains, there is a greater likelihood of new funds (with no embedded capital gains) entering the market. During the mid-1990s, a number of mutual fund families introduced "tax-managed mutual funds" that operated with reduced levels of capital-gain realizations. At the end of 1999, however, assets in "tax-managed" mutual funds represented just over one percent of the equity mutual fund marketplace, so these funds had not yet attracted a large share of the assets invested in the mutual fund sector. The growth of exchange-traded funds in the late 1990s, described in Poterba and Shoven (2002), are another way to reduce the tax burdens associated with holding a broad portfolio of securities.

4.3. Taxation and asset holding in tax-deferred accounts

One of the most dramatic developments in the structure of household portfolios during the last two decades has been the growing importance of assets held in defined-contribution pension plans. In the United States, the combined effects of growing regulatory burdens on defined-benefit pension plans, and increased worker mobility and the associated demand for portable pension arrangements, has led to a shift from defined-benefit to defined-contribution pension plans. Kruse (1995) and Gustman and Steinmeier (1992) discuss the reasons for these shifts. Samwick and Skinner (1998) explore how these changes in the structure of pension arrangements are influencing the nature of risk-bearing by households and the firms that offer various pension plans.

The fastest-growing type of defined-contribution plan in the United States is the so-called 401(k) pension plan, which permits workers to defer a share of their current earnings and the associated taxes while earning returns at the pre-tax rate. Poterba, Venti and Wise (2000) present summary information on the growth of these plans. Other tax-deferred methods of accumulation, such as Individual Retirement Accounts and 403(b) plans, have also grown in total assets, although their participant growth is slower than that for 401(k) plans. There has also been rapid growth in defined-contribution pension arrangements outside the United States. Personal Equity Plans in the United Kingdom, and Registered Retirement Saving Plans in Canada, for example, provide individuals with opportunities for tax deferral on investment income.

Assets held in tax-deferred accounts are a large and growing component of household net worth, and the portfolio allocation issues that arise in connection with these accounts have not been widely investigated. Most of the research to date on tax-deferred accounts, which is summarized in Bernheim (2001), has concentrated on the extent to which assets held in these accounts have "crowded out" other assets, or equivalently, on whether saving in tax-deferred accounts represents new saving. Very little research has considered the implications of tax-deferred accumulation opportunities for the structure of household portfolios, although this is an emerging topic that is attracting current research attention.

One aspect of the growth of defined-contribution-plan assets, with particularly important implications for studies of household portfolio behavior, is the growing

importance of the set of households that must make asset-allocation decisions in *both* taxable and tax-deferred accounts. Shoven (1999) outlined the "asset location problem", the problem of deciding whether to hold particular assets in a taxable account or in a tax-deferred retirement saving account. A number of subsequent studies, including Dammon, Spatt and Zhang (2000), Huang (2001), Poterba, Shoven and Sialm (2001), Shoven (1999) and Shoven and Sialm (1998, 2002), have considered various aspects of the asset-location problem. The key insight with respect to portfolio structure follows from an earlier literature on optimal corporate pension funding policy, such as Black (1980) and Tepper (1981). It is that investors should hold their highly taxed assets in their tax-deferred account and hold lightly taxed assets in their own taxable account. This advice is sometimes translated, loosely, as "stocks on own account, bonds in the tax-deferred account".

A key question in the recent asset-location literature concerns the identification of low-tax assets. If investors follow a buy-and-hold strategy with individual stocks in building their equity portfolio, the tax burden on their equity investments will be substantially smaller than if they purchase an average actively managed equity mutual fund. If they hold fixed-income assets by purchasing taxable corporate or government bonds, the total tax burden (considering both the implicit and explicit taxes on interest income) will be higher than if they hold tax-exempt bonds. It is possible, for some investment horizons and marginal-tax-rate configurations, for investors to find that holding actively managed equity mutual funds in their tax-deferred accounts is the preferred asset-location strategy.

Available evidence on asset allocation in tax-deferred accounts does not offer clear conclusions on the extent to which investors have "solved" the asset-location problem. Bodie and Crane (1997) present the most direct evidence to date on the extent to which investors recognize taxes in configuring their portfolios between taxable and tax-deferred accounts. They study an unusual data base on participants in TIAA-CREF, the defined-contribution pension system that covers most academics and other employees of college and universities in the United States. Their data combines a survey of TIAA-CREF participants with information on asset-allocation decisions within the retirement system. The results suggest that investors pursue similar asset-allocation strategies with respect to their taxable and tax-deferred accounts, and do not suggest clear understanding of the advantages to holding highly taxed assets in tax-deferred accounts.

Poterba and Wise (1998) present evidence on asset-allocation patterns in both IRAs and 401(k) plans. Roughly 46 percent of IRA and 401(k) assets are held in corporate equities, which are lightly taxed assets from the perspective of most taxable individuals. Bergstresser and Poterba (2001) use the Survey of Consumer Finances to explore asset holding by individual households, but the SCF data do not permit very precise inferences about portfolio holdings.

There are other issues associated with tax-deferred accounts and portfolio structure that have just begun to receive attention. One of the least studied but potentially important taxes for young households in the United States is the "tax" that is imposed

by the financial aid formula that colleges and universities use to determine student eligibility for scholarships and loans. Feldstein (1995) and Dick and Edlin (1997) argue that the tax rates implicit in the scholarship formula can be greater than those in the income-tax system for many households. The reason tax-deferred accounts are affected by this formula is that assets held in these accounts, which are deemed retirement saving, are not included in a household's net worth for the financial aid determination. Kim (1997) shows that households that are likely to face a higher marginal tax rate under the financial aid rules are more likely to hold assets in IRAs and 401(k) plans, and are likely to hold more assets in these accounts, than are similar households whose financial or family situations expose them to lower tax rates under the financial aid rules. Experience with the financial aid system is particularly relevant for understanding how means-tested transfer programs might affect saving by the elderly or other groups, and the impact of this tax on both the level and composition of this tax is worth further analysis.

4.4. Taxation and insurance products

Another class of assets that often receive specialized tax treatment, and that can represent an important share of household portfolios, is the class of insurance products. Many insurance policies, such as whole life insurance, deferred annuities, and variable annuities, combine both an insurance function and an investment component. For a variety of historical reasons, in many nations the "inside build up" on life insurance products is not taxed on accrual. In the United States, for example, if an individual purchases a deferred annuity policy when he is 45, but the annuity is not scheduled to begin until he reaches age 65, the capital income earned on his initial premium is not taxed until after the annuity payouts begin. Similarly, the income from variable annuities is not taxed until it is distributed to the investor. Whole life insurance policies offer a related opportunity to defer taxes on accruing interest and dividends.

The tax treatment of insurance products and the effect of tax rules on the demand for insurance is a potentially rich field for research, but it has attracted relatively little attention to date, especially from public-finance scholars interested in broad issues relating to taxation and capital accumulation. Several recent studies have described the tax treatment of insurance products and investigated the role of taxes in stimulating demand for these products. For example, Gentry and Milano (1998) explain how tax considerations affect the demand for variable annuities, which combine the investment flexibility of mutual funds with the favorable tax treatment of insurance products. Mitchell, Poterba, Warshawsky and Brown (1999) describe the tax treatment of life annuity products, but they do not attempt to measure the demand for these products.

The role of insurance products in wealth accumulation is smaller in the United States than in many other nations. Investigating the role of taxation in encouraging capital accumulation through insurance thus seems like an important issue for study in many nations.

4.5. The estate tax and portfolio structure

Some nations levy taxes on wealth, particularly when wealth-holders die and bequeath their assets to others. In the United States, the estate and gift tax currently raises one third of the estimated revenue from the capital-gains tax, but it is collected from a very small pool of decedents. Just over thirty thousand taxable estate tax returns are filed in a typical year. In 2000, decedents with net estates worth more than $675 000 were subject to estate tax. There is active political debate about raising this limit substantially, with elimination of the estate tax in the United States a serious possibility.

Studies of the estate tax typically recognize that there are a wide range of estate-tax avoidance strategies available to high-net-worth individuals. The extent to which these strategies are used to avoid taxes remains an open question, however. Cooper (1979) is the classic statement of the voluntary character of the estate tax. Scholes, Wolfson, Erickson, Maydew and Shevlin (2002) and Schmalbeck (2001) discuss a range of estate-tax planning techniques that high-net-worth households can use to reduce their tax liabilities. These include complex trust arrangements as well as simpler strategies such as donating assets to charity. The extent to which households avail themselves of estate-tax avoidance strategies is an open issue. Wolff (1996) presents evidence suggesting that the estate tax base in the United States is substantially eroded, while Poterba (2000b, 2001b) offers data suggesting that many households do not take advantage of even low-cost avoidance strategies.

The impact of estate taxes on wealth accumulation, and more specifically on the structure of household portfolios, has attracted substantial research attention in the last decade. These issues are difficult to address because of the very limited data on high-wealth households and their financial affairs in the public domain. Gale and Slemrod (2001) summarize much of the recent work on the economic impact of the estate tax. Researchers have studied how the estate tax affects charitable giving, for example in Joulfaian (1991), and a number of other behaviors. One issue of particular importance for portfolio structure concerns the interaction between the estate tax and the capital-gains tax. The current US tax code allows for the recipients of bequests to "step up their basis", which eliminates capital-gains tax liability on any appreciation of assets that are bequeathed. This creates a trade-off between capital-gains tax liability and estate-tax liability for wealthy households contemplating estate-tax avoidance strategies, such as lifetime giving, versus bequests. Auten and Joulfaian (1997) and Poterba (2001a) show that higher estate tax rates are associated with a smaller effect of capital-gains taxation on realization behavior.

One of the difficulties with modeling the behavioral impact of the estate tax involves the need to specify how it affects household budget sets. Poterba (2000b) suggests that the estate tax raises the required return on portfolio assets, although the magnitude of this effect depends critically on difficult-to-measure parameters involving estate-tax avoidance techniques. He argues that the expected value of the after-tax income that an individual will pay on his capital income, assuming for simplicity that all asset income comes in the form of interest, is $(1 - \tau_b)*r + p*(1 - \tau_e)*[1 + (1 - \tau_b)*r]$.

In this expression, p is the probability of death over the time period when the rate of return is r. This illustrates that the estate tax operates as a tax on capital, and that it raises the effective tax burden on capital income. Holtz-Eakin and Marples (2001) attempt to quantify the welfare cost of the estate tax under one set of assumptions about the nature of intergenerational transfers. Their study represents an important step toward modeling and evaluating the efficiency costs of this tax.

Moving from the impact of the estate tax on after-tax returns to a conclusion about how the tax affects wealth accumulation is complicated by the lack of agreement on why households save and leave bequests. Gale and Perozek (2001) note that the ultimate impact of the estate tax on saving decisions is ambiguous, and is likely to be quite sensitive to the reason households are saving and the structure of bequest motives.

The interaction between the estate tax and other tax provisions is potentially central to any analysis of the tax. Shoven and Wise (1998) observe that this interaction can lead to particularly high tax rates for households that save through tax-deferred retirement accounts. Bernheim (1986) makes the ingenious argument that the estate tax may actually reduce the revenue collected by the federal government, because one way to avoid the tax is to transfer assets from the older generation to younger generations well before death. If older taxpayers tend to be in higher tax brackets than the younger ones who receive asset transfers, then the process of estate-tax avoidance may result in lower taxes on capital income while the donor is alive. This is a complex argument and it has not yet been subject to enough empirical analysis to permit a judgement on its validity.

4.6. Stock options: another portfolio component

One aspect of portfolio behavior that has become increasingly important is the use of stock options as part of compensation packages. In a growing fraction of firms in the United States, employees receive wage and salary compensation as well as either a grant of corporate stock or options to purchase corporate stock. For employees at such firms, the value of stock options can become an important component of their total wealth. The interplay between decisions with respect to portfolio assets, and decisions about stock-option exercise, is an issue of growing importance that has yet to receive substantial attention from public-finance researchers. Huddart (1998) explains the tax consequence of various strategies with respect to the exercise of employee stock options. He also presents some evidence that many option holders do not exercise their options in a manner that would be consistent with tax-minimizing behavior. Huddart and Lang (1996) present a broader analysis of the factors that lead households to exercise employee stock options.

5. Taxation, risk-taking, and human capital

One issue that has received relatively little discussion in most analyses of the tax code and risk-taking is the impact of taxation on choice of occupation and more generally

on the *human capital* investments that individuals choose to make. In part, this reflects the difficulty of quantifying the dimensions along which individuals make choices regarding both their human-capital acquisition and their labor supply.

Public-finance researchers have long recognized that the tax treatment of returns on financial investments can affect the attractiveness of human-capital investments. Boskin (1975) and Heckman (1976) note that when capital-income taxes reduce the after-tax return on financial assets, they will induce individuals to acquire more human capital since the required rate of return on human capital will decline. Kaplow (1996) discusses related issues. Whether this insight carries over to a world with an income tax, rather than just a capital-income tax, depends on the structure of the tax. If the wage tax is proportional, and the cost of acquiring education is only foregone earnings, then the after-tax comparison between foregone earnings and the earnings increment associated with a human-capital investment will not be affected by the wage tax rate. With a progressive tax schedule, the returns to human-capital investment may be taxed at a higher rate than the tax rate at which the foregone earnings associated with training or education can be deducted. A progressive tax schedule therefore may discourage human-capital investment, although the wage progressivity effect can be offset by the rate of return effect described above.

These well-known results on income taxation and human-capital acquisition are derived in models of certainty. Eaton and Rosen (1980) move beyond this setting to explore how income taxes affect human-capital investment choices in a world of uncertainty. They find that the impact of earnings uncertainty on the distortionary costs of a wage tax depend on the structure of household preferences. The assumption of constant relative risk aversion, for example, can yield different results than constant absolute risk aversion. Hamilton (1987) presents further results on the structure of optimal taxes when the return to human capital is uncertain. He finds because individuals must bear idiosyncratic risk on their human-capital investments, they acquire less than the socially efficient level of human capital. An interest-income tax can encourage human-capital acquisition, and it may therefore be part of an optimal tax regime. Judd (1998) integrates the discussion of human-capital taxation in a world of uncertainty with recent insights about the long-run optimality of capital-income taxation. He concludes that how taxes on both financial assets and on wages affect human capital investment is likely to be sensitive to many detailed features of the utility function and of the political setting in which human-capital inputs, such as public schools, are provided.

The assumption that the foregone earnings associated with schooling or training are the primary cost of human-capital acquisition has been challenged in a number of recent studies. King and Rebelo (1990), Rebelo (1991) and Trostel (1993) find that when there are out-of-pocket school costs and tuition, then even a proportional wage tax can have a negative impact on investment in human capital. Lord and Rangazas (1998) present simulation findings that question this conclusion; they argue that the insurance effect of progressive wage taxation may be large enough to outweigh the decline in investment associated with higher tax burdens on returns than on the

inputs to human-capital investments. Heckman, Lochner and Taber (1998) present findings from a very detailed general-equilibrium model that incorporates endogenous human-capital acquisition as well as realistic descriptions of other aspects of household life-cycle behavior. They conclude that general-equilibrium responses in both the labor market and the capital market can substantially weaken conclusions about taxes and human-capital investment from partial equilibrium models. Most of their results suggest relatively limited effects of income taxation on human-capital investment.

The amount of human capital that individuals acquire is the primary focus of most models of taxes and human capital. It is also possible, however, that the tax system may affect the type of human capital that individuals acquire. There is little evidence, however, on the link between taxes and occupational choice, which is one measurable dimension of human capital type. One strand of research that does provide some insight in this issue concerns decisions about whether to enter self-employment or paid employment. Self-employment is often viewed as coincident with working in the entrepreneurial sector, a particularly high-risk sector of the labor market. Equating self-employment with founding of potentially high-growth firms is probably inappropriate; many self-employed individuals may simply be engaged in providing services such as painting or cleaning for which the growth opportunities are limited. However, those who do start new firms in various fields will be counted among the self-employed. Bruce (2000) and Schuetze (2000) present evidence that the combined level of income and payroll taxes on employed vs. self-employed workers affects the mix of workers in these two segments of the labor market.

Self-employment, particularly the type associated with starting new firms, may be affected by wage tax rates as well as capital-income tax rates. Since a substantial part of the labor income of self-employed individuals may be reinvested in the firm, it is possible that it will ultimately be taxed at the capital-gains tax rate rather than the labor-income tax rate, which may include both the payroll and the personal income tax rates. Thus the tax rate differential $(\tau_{cg} - \tau_{labor})$ may affect the supply of entrepreneurial talent. The level of self-employment may also be affected by the "demand" for entrepreneurs, or alternatively, by the supply of capital to start-up enterprises. This in turn is potentially affected by the relative tax treatment of the capital gains that investors in such start-ups might expect to receive, by comparison with the after-tax returns that they might expect on other investments. Thus, the tax rate differential $(\tau_{cg} - \tau_{int})$ may be a relevant factor in the supply of funds to start-up enterprises.

Poterba (1989) discusses these two channels for tax effects in more detail. The influence of taxation on the supply of funds may be greater at an earlier stage in the start-up process, when a prospective entrepreneur contacts the "informal" capital market to secure funding for a new enterprise. The so-called "angels", individuals who supply start-up capital, are likely to be sensitive to the capital-gains tax rate in making their capital supply decisions. Holtz-Eakin, Joulfaian and Rosen (1994a,b) investigate another aspect of the tax system, the estate tax, and its impact on the rate of new firm start-ups. They study the probability individuals who receive substantial inheritances will report income from self-employment in the years after they

receive their inheritance. They find that the probability that a self-employed person will remain self-employed rises if they receive a bequest, and that the chance that someone will enter self-employment is also an increasing function of intergenerational transfers. These findings are consistent with the view that self-employed individuals are capital-constrained, and that the supply of capital, which may be affected by tax rules, is an important determinant of the level of self-employment in the economy.

6. Conclusions and unresolved issues

The substantial theoretical and empirical literature on how taxation affects household portfolio behavior and risk-taking suggests a wide range of potential distortions. The empirical literature, while not offering universal support, generally suggests that taxation plays an important role in determining the set of assets households own, the amount that they invest in each of the available assets, when they sell assets, and the way risk is shared throughout the private economy. Measuring behavioral effects has proven easier than quantifying the welfare cost of behavioral distortions. There are few convincing estimates of the deadweight burdens associated either with distortions in portfolio structure or with changes in the timing and level of asset sales. Developing models of household portfolio behavior, and using these models to evaluate the welfare effects of tax policy, is an important research priority.

One of the challenges in studying taxation and household portfolio structure is the ever-changing nature of the tax and financial environment. Many studies summarized above assume, for example, that when individuals hold "stocks", they are taxed on their dividend income and realized capital gains, while if they hold "bonds", they are taxed on dividend income. This is an accurate depiction of the situation in which the household invests in stocks and bonds directly. But there are a wide range of ways for households to hold the risky streams that are associated with "stocks" and "bonds". A "stock" investor, for example, might buy a portfolio of individual stocks, or he might buy shares in a mutual fund, or he might buy futures on the S&P 500, or he might invest in an insurance product such as a variable annuity, or he might hold stocks or a mutual fund in an Individual Retirement Account. Each of these alternatives would have different tax consequences. The menu of ways to hold "stocks" is changing, even as this chapter is written. The rise of "exchange-traded funds" in the last half of the 1990s offers a new set of vehicles for holding common stocks. Tracking the effect of changes in the financial environment on household portfolio choices, and identifying the impact of taxation on these links, is a key ongoing subject of research.

Another challenge in analyzing taxation and portfolio behavior is that many of the households with substantial net worth receive sophisticated tax-planning advice. This advice may change the effective tax rates that these households face in important ways, yet it is difficult for academic researchers to incorporate such effects into empirical models. Henriques and Norris (1996) and Jacobs (1996) suggest that there are substantial opportunities to use sophisticated tax-planning strategies to avoid taxes.

Scholes, Wolfson, Erickson, Maydew and Shevlin (2002) also describe tax-avoidance opportunities, and they also outline some of the distortions associated with such activities. Quantifying the cost of tax-planning advice, and documenting the implicit taxes that high-income taxpayers face as they try to reduce their income and estate tax liabilities, is another important avenue for future study.

Besides modeling the changing financial system, and the impact of tax-planning advice on effective tax burdens, there are several other issues that call for further research. One concerns the dynamics of portfolio adjustment, and the factors that influence household decisions with regard to portfolio change. A significant literature in behavioral economics suggests that purely rational models of asset selection and asset management may not characterize household decision-making, and that households use rules of thumb and take time to adjust their behavior. One interesting issue is whether these behaviors apply to the high-net-worth households with substantial assets to invest. Another is how long it takes investors to respond to a substantial tax change, such as that in the United States in 1986. When tax systems are continuously subject to reform, the time to adjust can be an important determinant of the revenue effects and deadweight costs of the prevailing tax rules on portfolio income. Studies of household portfolio holdings are typically concerned with explaining the balance sheet "snapshot" at a given point in time, and in most cases they relate current holdings to current tax rules. With adjustment lags, however, the tax system in previous years can also have an important effect on current asset holdings.

Another issue that requires further research is the role of tax-code uncertainty in affecting portfolio choices. Most of the discussion in this chapter also assumes that investors know the tax code with certainty when they make investment decisions. Yet as Dickson (2000) and Sialm (2000) demonstrate, households face substantial uncertainty in the pattern of future tax rates. This uncertainty applies both with respect to the structure of marginal tax rates, and with respect to specific tax provisions that may apply to accumulated wealth. Capital-gains tax rules, and the rules that apply to withdrawals from retirement-saving accounts, are examples of such detailed provisions. While there has been some research on the impact of tax-code uncertainty on corporate investment, for example Hassett and Metcalf (1999), this issue has received less attention with respect to household decision-making. Recognizing "tax-code uncertainty" and incorporating it in models of household portfolio choice represents a useful avenue for future work.

A final issue that warrants attention is the effect of taxation on the overall level of asset prices. A number of studies cited in this chapter present evidence of some "tax capitalization", i.e., of a link between the level of tax rates and the value of particular assets. McGrattan and Prescott (2000) address a broader issue, and argue that changes in both dividend tax rates and in the composition of stock ownership in the United States have contributed substantially to the rise in the market value of US stocks during the 1990s. The tax changes in the United States and several other nations during the

last two decades have been large enough to admit the possibility of non-trivial effects on asset prices; further work could explore this relationship.

The impact of taxation on household portfolio behavior is an issue that already attracts attention in both applied tax-policy debates and in the academic disciplines of public economics and financial economics. But this issue is likely to become even more important prospectively. The aging of the "baby boom" generation in the United States, and the entry of large birth cohorts throughout the developed world into the age ranges in which asset accumulation becomes an important priority, suggests growing concern with issues associated with asset accumulation. The impact of taxes on asset accumulation, asset choice, and ultimately on the draw-down of wealth and the transfer of assets to the next generation is therefore likely to be a topic of growing interest and importance.

References

Agell, J., and P. Edin (1990), "Marginal taxes and the asset portfolios of Swedish households", Scandinavian Journal of Economics 92:47–64.

Allen, F., and R. Michaely (1995), "Dividend policy", in: R. Jarrow, V. Maksimovic and W. Ziemba, eds., Handbook of Operations Research and Management Science: Finance (Amsterdam, North Holland) pp. 793–834.

Allen, F., A. Bernardo and I. Welch (2000), "A theory of dividends based on tax clienteles", Journal of Finance 55:2499–2536.

Amoako-Adu, B., M. Rashid and M. Stebbins (1992), "Capital gains tax and equity values: empirical test of stock price reaction to the introduction and reduction of capital gains tax exemption", Journal of Banking and Finance 16:275–287.

Arnott, R., A. Berkin and J. Ye (2000), "How well have taxable investors been served in the 1980s and 1990s?", Journal of Portfolio Management 26(Summer):84–93.

Atkinson, A.B., and A. Sandmo (1980), "Welfare implications of the taxation of savings", Economic Journal 90:529–549.

Auerbach, A.J. (1983), "Stockholder tax rates and firm attributes", Journal of Public Economics 21:107–127.

Auerbach, A.J. (1988), "Capital gains taxation in the United States: realizations, revenue, and rhetoric", Brookings Papers on Economic Activity 2:595–631.

Auerbach, A.J. (1991), "Retrospective capital gains taxation", American Economic Review 81:167–178.

Auerbach, A.J. (1992), "On the design and reform of capital gains taxation", American Economic Review 82(May):263–267.

Auerbach, A.J. (2001), "Taxation and corporate financial policy", in: A.J. Auerbach and M.S. Feldstein, eds., Handbook of Public Economics, Vol. 3 (Elsevier, Amsterdam) pp. 1251–1292.

Auerbach, A.J., and M.A. King (1982), "Corporate financial policy with personal and institutional investors", Journal of Public Economics 17:259–285.

Auerbach, A.J., and M.A. King (1983), "Taxation, portfolio choice, and debt-equity ratios: a general equilibrium model", Quarterly Journal of Economics 98:587–609.

Auerbach, A.J., and J. Siegel (2000), "Capital gains realizations of the rich and sophisticated", American Economic Review 90(May):276–282.

Auerbach, A.J., L. Burman and J. Siegel (2000), "Capital gains taxation and tax avoidance: new evidence from panel data", in: J. Slemrod, ed., Does Atlas Shrug? (Russell Sage Foundation, New York) pp. 355–388.

Auten, G., and C.T. Clotfelter (1982), "Permanent vs. transitory tax effects and the realization of capital gains", Quarterly Journal of Economics 97:613–632.

Auten, G., and D. Joulfaian (1997), "Bequest taxes and capital gains realizations", Mimeo (U.S. Treasury Department, Office of Tax Analysis, Washington, D.C.).

Auten, G., and D. Joulfaian (1999), "How income taxes affect capital gains realizations: evidence from a long panel", Mimeo (U.S. Treasury Department, Office of Tax Analysis, Washington, D.C.).

Ayers, B., C. Cloyd and J. Robinson (2000), "Capitalization of shareholder taxes in stock prices: evidence from the Revenue Reconciliation Act of 1993", Mimeo (University of Texas-Austin Graduate School of Business).

Badrinath, S., and W. Lewellen (1991), "Evidence on tax-motivated securities trading behavior", Journal of Finance 46:369–382.

Bailey, M. (1969), "Capital gains and income taxation", in: A. Harberger and M. Bailey, eds., The Taxation of Income from Capital (The Brookings Institution, Washington).

Balcer, Y., and K. Judd (1987), "Effects of capital gains taxation on life-cycle investment and portfolio management", Journal of Finance 42:743–761.

Bali, R., and G. Hite (1998), "Ex-dividend day stock price behavior: discreteness or tax-induced clienteles?", Journal of Financial Economics 47:127–159.

Barclay, M. (1987), "Dividends, taxes, and common stock prices: the ex-dividend day behavior of common stock prices before the income tax", Journal of Financial Economics 19:31–44.

Barclay, M., N. Pearson and M. Weisbach (1998), "Open end mutual funds and capital gains taxes", Journal of Financial Economics 49:4–43.

Barlow, R., H. Brazer and J. Morgan (1966), Economic Behavior of the Affluent (Brookings Institution, Washington).

Basak, S., and B. Croitoru (2001), "Nonlinear taxation, tax arbitrage, and equilibrium asset prices", Journal of Mathematical Economics 35:347–382.

Basak, S., and M. Gallmeyer (1998), "Capital market equilibrium with differential taxation", Mimeo (Wharton School, University of Pennsylvania).

Bell, L., and T. Jenkinson (2000), "New evidence of the impact of dividend taxation and on the identity of the marginal investor", Mimeo (Department of Economics, Oxford University).

Bergstresser, D., and J.M. Poterba (2001), "Household asset location decisions: evidence from the Survey of Consumer Finance", Mimeo (MIT Department of Economics).

Bergstresser, D., and J.M. Poterba (2002), "Do after-tax returns affect mutual fund inflows?", Journal of Financial Economics (forthcoming).

Berkovec, J., and D. Fullerton (1992), "A general equilibrium model of housing, taxes, and portfolio choice", Journal of Political Economy 100:390–429.

Bernheim, B.D. (1986), "Does the estate tax raise revenue?", in: L. Summers, ed., Tax Policy and the Economy, Vol. 1 (MIT Press, Cambridge) pp. 113–138.

Bernheim, B.D. (2001), "Taxation and saving", in: A.J. Auerbach and M.S. Feldstein, eds., Handbook of Public Economics, Vol. 3 (Elsevier, Amsterdam) pp. 1173–1249.

Bertaut, C., and M. Starr-McCluer (2001), "Household portfolios in the United States", in: L. Guiso, M. Haliassos and T. Jappelli, eds., Household Portfolios (MIT Press, Cambridge) in press.

Bhabra, H., U. Dhillon and G. Ramirez (1999), "A November effect? Revisiting the tax loss selling hypothesis", Financial Management 28:5–15.

Bhardwaj, R., and L. Brooks (1999), "Further evidence on dividend yields and the ex-dividend day stock price effect", Journal of Financial Research 22(Winter):503–514.

Black, F. (1976), "The dividend puzzle", Journal of Portfolio Management 2(Winter):5–8.

Black, F. (1980), "The tax consequences of long-run pension policy", Financial Analysts Journal (July/August), 21–28.

Blouin, J., J. Raedy and D. Shackelford (2000), "Capital gains holding periods and equity trading: evidence from the 1998 tax act", NBER Working Paper 8011 (NBER).

Blume, M., J. Crockett and I. Friend (1974), "Stockownership in the United States: characteristics and trends", Survey Current Business 54(November):16–40.
Bodie, Z., and D. Crane (1997), "Personal investing: advice, theory, and evidence", Financial Analysts Journal 53(November/December):13–23.
Bolster, P., L. Lindsey and A. Mitrusi (1989), "Tax induced trading: the effect of the 1986 Tax Reform Act on stock market activity", Journal of Finance 44:327–344.
Booth, L., and D. Johnston (1984), "The ex-dividend day behaviour of Canadian stock prices: tax changes and clientele effects", Journal of Finance 39:457–476.
Boskin, M. (1975), "Notes on the tax treatment of human capital", in: U.S. Treasury Department Conference on Tax Research (U.S. Treasury Department, Washington).
Bossons, J. (1972), "An economic overview of the tax reforms", in: Proceedings of the 23rd Tax Conference, Toronto (Canadian Tax Foundation) pp. 45–67.
Bradford, D. (1995), "Fixing realization accounting: symmetry, consistency, and correctness in the taxation of financial instruments", Tax Law Review 50:731–786.
Brennan, M. (1970), "Taxes, market valuation, and corporate financial policy", National Tax Journal 23:417–427.
Brickley, J., S. Manaster and J.S. Schallheim (1991), "The tax-timing option and discounts on closed end investment companies", Journal of Business 64:287–312.
Bruce, D. (2000), "Effects of the United States tax system on transitions into self employment", Labour Economics 7:545–574.
Bulow, J.I., and L.H. Summers (1984), "The taxation of risky assets", Journal of Political Economy 92:20–39.
Burman, L. (1999), The Labyrinth of Capital Gains Tax Policy (Brookings Institution, Washington).
Burman, L., and W. Randolph (1994), "Measuring permanent responses to capital gains tax changes in panel data", American Economic Review 84:794–809.
Burman, L., S. Wallace and D. Weiner (1997), "How capital gains taxes distort homeowners decisions", in: Proceedings of the 89th Annual Conference on Taxation (National Tax Assocation, Columbus, OH) pp. 382–390.
Butters, J., L. Thompson and L. Bollinger (1953), Effects of Taxation: Investment by Individuals (Graduate School of Business Administration, Harvard University, Cambridge, MA).
Campbell, J.Y., and K.A. Froot (1995), "Securities transaction taxes: what about international experiences and migrating markets", in: S. Hammond, ed., Securities Transaction Taxes: False Hopes and Unintended Consequences (Catalyst Institute, Chicago) pp. 110–142.
Chalmers, J. (1998), "Default risk cannot explain the municipal puzzle: Evidence from municipal bonds that are secured by U.S. Treasury obligations", Review of Financial Studies 11(Summer):281–308.
Chaplinsky, S., and H. Seyhun (1990), "Dividends and taxes: Evidence on tax-reduction strategies", Journal of Business 63:239–260.
Chay, J., D. Choi and J. Pontiff (2000), "Market valuation of tax-timing options: Evidence from capital gains distributions", Mimeo (University of Washington Department of Finance).
Constantinides, G. (1984), "Optimal stock trading with personal taxes: implications for prices and the abnormal January returns", Journal of Financial Economics 13:65–89.
Constantinides, G., and M.S. Scholes (1980), "Optimal liquidation of assets in the presence of personal taxes: implications for asset pricing", Journal of Finance 35:439–449.
Cooper, G. (1979), A Voluntary Tax? New Perspectives on Sophisticated Estate Tax Avoidance (The Brookings Institution, Washington).
Dammon, R., and R. Green (1987), "Tax arbitrage and the existence of equilibrium prices for financial assets", Journal of Finance 42:1143–1166.
Dammon, R., and C. Spatt (1996), "The optimal trading and pricing of securities with asymmetric capital gains taxes and transactions costs", Review of Financial Studies 9:921–952.
Dammon, R., C. Spatt and H. Zhang (2000), "Optimal asset location and allocation with taxable

and tax-deferred investing", Mimeo (Carnegie Mellon University, Graduate School of Industrial Administration).
Dammon, R., C. Spatt and H. Zhang (2001), "Optimal consumption and investment with capital gains taxes", Review of Financial Studies 14:583–616.
Dhaliwal, D., M. Erickson and R. Trezevant (1999), "A test of the theory of dividend clienteles", National Tax Journal 52:179–194.
Dick, A., and A.S. Edlin (1997), "The implicit taxes from college financial aid", Journal of Public Economics 65:295–322.
Dickson, J. (2000), "Pension taxation and tax code risk", in: W. Gale, J. Shoven and M. Warshawsky, eds., Public Policies and Private Pensions (Brookings Institution, Washington) in press.
Dickson, J., and J.B. Shoven (1994), "A stock index mutual fund without net capital gain realizations", NBER Working Paper 4717 (NBER, Cambridge, MA).
Dickson, J., and J.B. Shoven (1995), "Taxation and funds: an investor perspective", in: J. Poterba, ed., Tax Policy and the Economy, Vol. 9 (MIT Press, Cambridge, MA) pp. 151–171.
Dickson, J., J.B. Shoven and C. Sialm (2000), "Tax externalities of equity mutual funds", National Tax Journal 53(September, Part 2):607–628.
Domar, E.D., and R.A. Musgrave (1944), "Proportional income taxation and risk-taking", Quarterly Journal Economics 58:388–422.
Dyl, E. (1977), "Capital gains taxation and year-end stock market behavior", Journal of Finance 32:165–175.
Eades, K., P. Hess and E. Kim (1984), "On interpreting security returns during the ex-dividend period", Journal of Financial Economics 13:3–34.
Eades, K., P. Hess and E. Kim (1994), "Time series variation in dividend pricing", Journal of Finance 49:1617–1638.
Eaton, J., and H.S. Rosen (1980), "Taxation, human capital, and uncertainty", American Economic Review 70:705–715.
Eichner, M., and T. Sinai (2000), "Capital gains tax realizations and tax rates: new evidence from time series", National Tax Journal 53(September, part 2):663–682.
Elton, E., and M. Gruber (1970), "Marginal stockholders' tax rates and the clientele effect", Review of Economics and Statistics 52:68–74.
Elton, E., and M. Gruber (1978), "Taxes and portfolio composition", Journal of Financial Economics 6:399–410.
Feenberg, D.R. (1981), "Does the investment interest limitation explain the existence of dividends?", Journal of Financial Economics 9:265–269.
Feenberg, D.R., and E. Coutts (1993), "An introduction to the TAXSIM model", Journal of Policy Analysis and Management 12:189–194.
Feenberg, D.R., and J.M. Poterba (1991), "Which households own municipal bonds? Evidence from tax returns", National Tax Journal 44(December):93–103.
Feldstein, M. (1976), "Personal taxation and portfolio composition: an econometric analysis", Econometrica 44:631–649.
Feldstein, M. (1978), "The welfare cost of capital income taxation", Journal of Political Economy 86:S29–S52.
Feldstein, M. (1995), "College scholarship rules and private saving", American Economic Review 85:552–566.
Feldstein, M., J. Slemrod and S. Yitzhaki (1980), "The effects of taxation on the selling and switching of common stock", Quarterly Journal of Economics 94:777–791.
Fortune, P. (1988), "Municipal bond yields: whose tax rates matter?", National Tax Journal 41:219–233.
Frank, M., and R. Jagannathan (1998), "Why do stock prices drop by less than the value of the dividend? Evidence from a country without taxes", Journal of Financial Economics 47:161–188.
Gale, W., and M. Perozek (2001), "Does the estate tax affect saving?", in: W. Gale, J. Hines and J. Slemrod, eds., Rethinking Estate and Gift Taxation (Brookings Institution, Washington).

Gale, W., and J. Slemrod (2001), "Rethinking estate and gift taxes: overview", in: W. Gale and J. Slemrod, eds., Rethinking Estate and Gift Taxation (Brookings Institution, Washington) pp. 1–64.
Gentry, W., and J. Milano (1998), "Taxation and investment in annuities", Working Paper 6526 (National Bureau of Economic Research, Cambridge, MA).
Ghee, W., and W. Reichenstein (1996), "The after-tax returns from different saving vehicles", Financial Analysts Journal (July/August) 62–72.
Gordon, R.H. (1985), "Taxation of corporate capital income: tax revenues versus tax distortions", Quarterly Journal of Economics 100:1–27.
Gordon, R.H., and D. Bradford (1980), "Taxation and the stock market value of capital gains and dividends: theory and empirical results", Journal of Public Economics 14:109–136.
Gravelle, J.G. (1994), The Economic Effects of Taxing Capital Income (MIT Press, Cambridge, MA).
Green, R. (1993), "A simple model of the taxable and tax-exempt yield curves", Review of Financial Studies 6:233–264.
Green, R., and B. Odegaard (1997), "Are there tax effects in the relative pricing of U.S. government bonds?", Journal of Finance 52:609–633.
Green, R., and K. Rydqvist (1999), "Ex-day behavior with dividend preference and limitations to short-term arbitrage: the case of Swedish lottery bonds", Journal of Financial Economics 53:145–187.
Grinblatt, M., and M. Keloharju (2000), "Tax-loss trading and wash sales", Mimeo (Anderson School of Management, UCLA).
Grinblatt, M., and M. Keloharju (2001), "What makes investors trade?", Journal of Finance 56:586–616.
Grinblatt, M., and T. Moskowitz (2000), "The cross section of expected returns and its relation to past returns: new evidence", Mimeo (UCLA Department of Finance).
Gruber, M. (1996), "Another puzzle: the growth in actively managed mutual funds", Journal of Finance 51:783–810.
Guiso, L., M. Haliassos and T. Jappelli (2001), Household Portfolios (MIT Press, Cambridge) in press.
Gustman, A.L., and T.L. Steinmeier (1992), "The stampede toward defined contribution pension plans: fact or fiction?", Industrial Relations 31:361–369.
Haliassos, M., and C. Bertaut (1995), "Why do so few hold stocks", Economic Journal 105:1110–1129.
Hamilton, J.H. (1987), "Optimal wage and income taxation with wage uncertainty", International Economic Review 28:373–388.
Hassett, K.A., and G.E. Metcalf (1999), "Investment with uncertain tax policy: does random tax policy discourage investment?", Economic Journal 109:372–393.
Heckman, J.J. (1976), "A life-cycle model of earnings, learning, and consumption", Journal of Political Economy 84:S11–44.
Heckman, J.J., L. Lochner and C. Taber (1998), "Tax policy and human capital formation", American Economic Review 88(May):293–297.
Henriques, D., and F. Norris (1996), "Wealthy, helped by Wall Street, find new ways to escape tax on profits", New York Times (December 1) 1.
Hochguertel, S., R. Alesie and A. van Soest (1997), "Saving accounts versus stocks and bonds in household portfolio allocation", Scandinavian Journal of Economics 99:81–97.
Holtz-Eakin, D., and D. Marples (2001), "Distortion costs of taxing wealth accumulation: income versus estate taxes", Working Paper 8261 (National Bureau of Economic Research, Cambridge, MA).
Holtz-Eakin, D., D. Joulfaian and H.S. Rosen (1994a), "Sticking it out: entrepreneurial survival and liquidity constraints", Journal of Political Economy 102:53–75.
Holtz-Eakin, D., D. Joulfaian and H.S. Rosen (1994b), "Entrepreneurial decisions and liquidity constraints", RAND Journal of Economics 23:334–347.
Huang, J. (2001), "Taxable or tax-deferred account? Portfolio decisions with multiple investment goals", Mimeo (MIT Sloan School of Management).
Hubbard, R.G. (1985), "Personal taxation, pension wealth, and portfolio composition", Review of Economics and Statistics 67:53–60.

Hubbard, R.G. (1995), "Securities transaction taxes: can they raise revenue?", in: S. Hammond, ed., Securities Transaction Taxes: False Hopes and Unintended Consequences (Catalyst Institute, Chicago) pp. 27–57.

Huddart, S. (1998), "Tax planning and the exercise of employee stock options", Contemporary Accounting Research 15(Summer):203–216.

Huddart, S., and M. Lang (1996), "Employee stock option exercises: an empirical analysis", Journal of Accounting and Economics 21:5–43.

Huddart, S., and V. Narayanan (2000), "An empirical examination of tax factors and mutual fund stock sales", Mimeo (Penn State University).

Ibbotson Associates (1996), Stocks, Bills, Bonds, and Inflation (Ibbotson Associates, Chicago).

Investment Company Institute (1999), Equity Ownership in America (Investment Company Institute and Securities Industry Association, Washington).

Ioannides, Y. (1989), "Housing, other real estate, and wealth portfolios", Regional Science and Urban Economics 19:261–280.

Jacobs, N. (1996), "Tax-efficient investing: reduce tax drag, improve asset growth", Trusts and Estates (June) 25–33.

Jeffrey, R., and R. Arnott (1993), "Is your alpha big enough to cover its taxes?", Journal of Portfolio Management (Spring) 15–25.

Joulfaian, D. (1991), "Charitable bequests and estate taxes", National Tax Journal 44:169–180.

Judd, K.L. (1998), "Taxes, uncertainty, and human capital", American Economic Review 88(May):289–292.

Kalay, A. (1982), "The ex-dividend day behavior of stock prices: a re-examination of the clientele effect", Journal of Finance 37:1059–1070.

Kaplow, L. (1994), "Taxation and risk-taking: a general equilibrium perspective", National Tax Journal 47:789–798.

Kaplow, L. (1996), "On the divergence between "ideal" and conventional income-tax treatment of human capital", American Economic Review 86(May):347–352.

Karpoff, J., and R. Walkling (1988), "Short-term trading around ex-dividend days: additional evidence", Journal of Financial Economics 21:291–298.

Karpoff, J., and R. Walkling (1990), "Dividend capture in NASDAQ stocks", Journal of Financial Economics 28:39–66.

Kennickell, A., M. Starr-McCluer and B. Surette (2000), "Changes in U.S. family finances at the end of the 1990s: results from the 1998 Survey of Consumer Finances", Federal Reserve Bulletin 86(January):1–29.

Khorana, A., and H. Servaes (1999), "The determinants of mutual fund starts", Review of Financial Studies 12:1043–1073.

Kiefer, D. (1990), "Lock-in effect within a simple model of corporate stock trading", National Tax Journal 43:75–94.

Kim, T. (1997), "College financial aid and family saving", Ph.D. dissertation (MIT Department of Economics, Cambridge, MA).

King, M.A., and J. Leape (1998), "Wealth and portfolio composition: theory and evidence", Journal of Public Economics 69:155–193.

King, R.G., and S. Rebelo (1990), "Public policy and economic growth: developing neoclassical implications", Journal of Political Economy 98:S126–S150.

Klein, P. (1999), "The capital gain lock-in effect and equilibrium returns", Journal of Public Economics 71(March):355–378.

Kochin, L., and R. Parks (1988), "Was the tax-exempt bond market inefficient or were future expected tax rates negative?", Journal of Finance 53:913–931.

Koski, J. (1996), "A microstructure analysis of ex-dividend stock price behavior before and after the 1984 and 1986 Tax Reform Acts", Journal of Business 69:313–338.

Koski, J., and J. Scruggs (1998), "Who trades around the ex-dividend day? Evidence from NYSE audit file data", Financial Management 27(Autumn):58–72.
Kovenock, D., and M. Rothschild (1987), "Notes on the effect of capital gains taxation on non-Austrian assets", in: A. Razin and E. Sadka, eds., Economic Policy in Theory and Practice (St. Martin's Press, New York) pp. 309–339.
Kraft, A., and I. Weiss (1998), "Tax planning by mutual funds: evidence from changes in the capital gains tax rate", Mimeo (University of Chicago Graduate School of Business).
Kruse, D.L. (1995), "Pension substitution in the 1980s: why the shift toward defined contribution?", Industrial Relations 34:218–241.
Lakonishok, J., and T. Vermaelen (1983), "Tax reform and ex-dividend behavior", Journal of Finance 38:1157–1179.
Landsman, W., and D. Shackelford (1995), "The lock-in effect of capital gains taxes: evidence from the RJR Nabisco leveraged buyout", National Tax Journal 58:245–260.
Lang, M., and D. Shackelford (2000), "Capitalization of capital gains taxes: evidence from stock price reactions to the 1997 rate reduction", Journal of Public Economics 76:69–85.
Leape, J. (1987), "Taxes and transaction costs in asset market equilibrium", Journal of Public Economics 33:1–20.
Leland, H.E. (2000), "Optimal portfolio management with transactions costs and capital gains taxes", Mimeo (Haas School of Business, University of California-Berkeley).
Lewellen, W., K. Stanley, R. Lease and G. Schlarbaum (1978), "Some direct evidence on the dividend clientele phenomenon", Journal of Finance 33:163–195.
Litzenberger, R., and K. Ramaswamy (1979), "The effect of personal taxes and dividends on capital asset prices: theory and empirical evidence", Journal of Financial Economics 7:163–195.
Litzenberger, R., and K. Ramaswamy (1980), "Dividends, short-selling restrictions, tax-induced investor clienteles, and market equilibrium", Journal of Finance 35:469–482.
Long, J. (1977), "Efficient portfolio choice with differential taxation of dividends and capital gains", Journal of Financial Economics 5:25–53.
Long, J. (1978), "The market valuation of cash dividends: a case to consider", Journal of Financial Economics 6:235–264.
Lord, W., and P. Rangazas (1998), "Capital accumulation and taxation in a general equilibrium model with risky human capital", Journal of Macroeconomics 20:509–531.
Lybeck, J. (1991), "On political risk – the turnover tax on the Swedish money and bond markets, or how to kill a market without really trying", in: S.J. Khoury, ed., Recent Developments in International Banking and Finance (North Holland, Amsterdam).
MacKie-Mason, J.K. (1990), "Some nonlinear tax effects on asset values and investment decisions under uncertainty", Journal of Public Economics 42:301–328.
Maki, D.M. (1996), "Portfolio shuffling and tax reform", National Tax Journal 49:317–330.
Mariger, R. (1995), "Taxes, capital gains realizations, and revenues: a critical review and some new results", National Tax Journal 48:447–462.
McDonald, R. (1983), "Government debt and private leverage", Journal of Public Economics 22:303–325.
McGrattan, E.R., and E.C. Prescott (2000), "Is the stock market overvalued?", NBER Working Paper 8077 (NBER, Cambridge, MA).
Meehan, J., D. Yoo and G. Fong (1995), "Taxable asset allocation with varying market risk premiums", Journal of Portfolio Management (Fall) 79–87.
Michaely, R. (1991), "Ex-dividend day stock price behavior: the case of the 1986 Tax Reform Act", Journal of Finance 46:845–860.
Michaely, R., and J.-L. Vila (1996), "Trading volume with private valuations: evidence from the ex-dividend day", Review of Financial Studies 9:471–509.
Miller, M.H. (1977), "Debt and taxes", Journal of Finance 32:261–275.
Miller, M.H., and M.S. Scholes (1978), "Dividends and taxes", Journal of Financial Economics 6:333–364.

Miller, M.H., and M.S. Scholes (1982), "Dividends and taxes: some empirical evidence", Journal of Political Economy 90:1118–1141.

Mintz, J., and M. Smart (2001), "Tax-exempt investors and corporate capital structure", Journal of Public Economics (forthcoming).

Mitchell, O.S., J.M. Poterba, M. Warshawsky and J. Brown (1999), "New evidence on the money's worth of individual annuities", American Economic Review 89(December):1299–1318.

Mitrusi, A., and J.M. Poterba (2001), "The changing importance of income and payroll taxes on U.S. families", in: J. Poterba, ed., Tax Policy and the Economy, Vol. 15 (MIT Press, Cambridge) pp. 95–120.

Morgan, G., and S. Thomas (1998), "Taxes, dividend yields, and returns in the U.K. equity market", Journal of Banking and Finance 22:405–423.

Naranjo, A., M. Nimalendran and M. Ryngaert (1998), "Stock returns, dividend yields, and taxes", Journal of Finance 53:2029–2057.

Odean, T. (1998), "Are investors reluctant to realize their losses?", Journal of Finance 53:1775–1798.

OECD (1994), Taxation and Household Saving (Organization for Economic Cooperation and Development, Paris).

Peterson, P., D. Peterson and J. Ang (1985), "Direct evidence on the marginal rate of taxation on dividend income", Journal of Financial Economics 14:267–282.

Petit, R. (1977), "Taxes, transactions costs, and the clientele effect of dividends", Journal of Financial Economics 5:419–436.

Poterba, J.M. (1986a), "The market value of cash dividends: the Citizens Utilities case reconsidered", Journal of Financial Economics 15:395–405.

Poterba, J.M. (1986b), "Explaining the yield spread between taxable and tax-exempt bonds: the role of expected tax policy", in: H. Rosen, ed., Studies in State and Local Public Finance (University of Chicago Press, Chicago) pp. 5–49.

Poterba, J.M. (1987), "How burdensome are capital gains taxes?", Journal of Public Economics 33:153–172.

Poterba, J.M. (1989), "Capital gains tax policy toward entrepreneurship", National Tax Journal 42: 375–390.

Poterba, J.M. (2000a), "Stock market wealth and consumption", Journal of Economic Perspectives 14(Spring):99–118.

Poterba, J.M. (2000b), "The estate tax and after-tax investor returns", in: J. Slemrod, ed., Does Atlas Shrug? (Russell Sage Foundation, New York) pp. 329–349.

Poterba, J.M. (2001a), "Taxation and portfolio structure: issues and implications", in: L. Guiso, M. Haliassos and T. Jappelli, eds., Household Portfolios (MIT Press, Cambridge) forthcoming.

Poterba, J.M. (2001b), "Incentive effects of estate and gift taxes in the United States", Journal of Public Economics 79:237–264.

Poterba, J.M., and A. Samwick (2002), "Taxation and household portfolio composition: evidence from the 1980s and 1990s", Journal of Public Economics (forthcoming).

Poterba, J.M., and J.B. Shoven (2002), "Exchange traded funds: new investment options for taxable investors", American Economic Review 92(May), forthcoming.

Poterba, J.M., and L.H. Summers (1984), "New evidence that taxes affect the valuation of dividends", Journal of Finance 39:1397–1415.

Poterba, J.M., and L.H. Summers (1985), "The economic effects of dividend taxation", in: E. Altmann and M. Subrahmanyam, eds., Recent Advances in Corporate Finance (Dow-Jones Irwin, Homewood, IL) pp. 227–284.

Poterba, J.M., and S. Weisbenner (2001a), "Taxing estates or unrealized capital gains at death, in: W. Gale, J. Hines and J. Slemrod, eds., Rethinking Estate and Gift Taxation (Brookings Institution, Washington) pp. 422–456.

Poterba, J.M., and S. Weisbenner (2001b), "Capital gains tax rules, tax loss trading, and turn of the year returns", Journal of Finance 56:353–368.

Poterba, J.M., and D.A. Wise (1998), "Individual financial decisions in retirement saving plans and the provision of resources for retirement", in: M. Feldstein, ed., Privatizing Social Security (University of Chicago Press, Chicago) pp. 363–393.
Poterba, J.M., S.F. Venti and D.A. Wise (2000), "Saver behavior and 401(k) wealth", American Economic Review 90(May):297–302.
Poterba, J.M., J.B. Shoven and C. Sialm (2001), "Asset location for retirement savers", in: W. Gale and J. Shoven, eds., Public Policies and Private Pensions (Brookings Institution, Washington) forthcoming.
Prisman, E., G. Roberts and Y. Tian (1996), "Optimal bond trading and the tax-timing option in Canada", Journal of Banking and Finance 20:1351–1363.
Protopapadakis, A. (1983), "Some indirect evidence on effective capital gains tax rates", Journal of Business 56:127–138.
Rebelo, S. (1991), "Long run policy analysis and long-run growth", Journal of Political Economy 99:500–521.
Reese, W. (1998), "Capital gains taxation and stock market activity: evidence from IPOs", Journal of Finance 53:1799–1819.
Ross, S. (1985), "Debt and taxes and uncertainty", Journal of Finance 40:637–657.
Samwick, A. (2000), "Portfolio responses to taxation: evidence from the end of the rainbow", in: J. Slemrod, ed., Does Atlas Shrug? (Russell Sage Foundation, New York) pp. 289–323.
Samwick, A., and J. Skinner (1998), "How will defined contribution pension plans affect retirement income?" NBER Working Paper 6645 (NBER, Cambridge, MA).
Sandmo, A. (1977), "Portfolio theory, asset demand, and taxation: comparative statics with many assets", Review of Economic Studies 44:369–379.
Sandmo, A. (1985), "The effects of taxation on savings and risk-taking", in: A. Auerbach and M. Feldstein, eds., Handbook of Public Economics, Vol. 1 (North Holland, Amsterdam) pp. 265–309.
Schmalbeck, D. (2001), "Avoiding federal wealth transfer taxes", in: W. Gale and J. Slemrod, eds., Rethinking Estate and Gift Taxation (Brookings Institution, Washington) pp. 113–158.
Scholes, M.S., M.A. Wolfson, M. Erickson, E. Maydew and T. Shevlin (2002), Taxes and Business Strategy: A Planning Approach, Second Edition (Prentice Hall, Upper Saddle River, NJ).
Scholz, J.K. (1992), "A direct examination of the dividend clientele hypothesis", Journal of Public Economics 49:261–285.
Scholz, J.K. (1994), "Tax progressivity and household portfolios: descriptive evidence from the surveys of consumer finances", in: J. Slemrod, ed., Tax Progressivity and Income Inequality (Cambridge University Press, New York) pp. 219–274.
Schuetze, H. (2000), "Taxes, economic conditions, and recent trends in male self-employment: a Canada–U.S. comparison", Labour Economics 7:507–544.
Schwert, G., and P. Seguin (1995), "Securities transaction taxes: an overview of costs, benefits, and unresolved questions", in: S. Hammonds, ed., Securities Transaction Taxes: False Hopes and Unintended Consequences (Catalyst Institute, Chicago) pp. 1–26.
Seida, J., and W. Wempe (2000), "Do capital gain tax rate increases affect individual investors' trading decisions?", Journal of Accounting and Economics 30:33–57.
Seyhun, H., and D. Skinner (1994), "How do taxes affect investors' stock market realizations? Evidence from tax-return panel data", Journal of Business 67:231–262.
Shackelford, D. (2000a), "The tax environment facing the wealthy", in: J. Slemrod, ed., Does Atlas Shrug? (Russell Sage Foundation, New York) pp. 114–138.
Shackelford, D. (2000b), "Stock market reaction to capital gains tax changes: empirical evidence from the 1997 and 1998 tax acts", in: J. Poterba, ed., Tax Policy and the Economy, Vol. 14 (MIT Press, Cambridge) pp. 67–92.
Shackelford, D., and R. Verrecchia (1999), "Intertemporal tax discontinuities", NBER Working Paper 7451 (NBER, Cambridge, MA).
Shefrin, H., and M. Statman (1985), "The disposition to sell winners too early and ride losers too long: theory and evidence", Journal of Finance 40:777–790.

Shoven, J.B. (1999), "The location and allocation of assets in pension and conventional savings accounts", NBER Working Paper 7007 (NBER, Cambridge, MA).
Shoven, J.B., and C. Sialm (1998), "Long run asset allocation for retirement saving", Journal of Private Portfolio Management 1(2):13–26.
Shoven, J.B., and C. Sialm (2002), "Asset location in tax-deferred and conventional savings accounts", Working Paper 7192 (National Bureau of Economic Research, Cambridge, MA).
Shoven, J.B., and D.A. Wise (1998), "The taxation of pensions: a shelter can become a trap", in: D. Wise, ed., Frontiers in the Economics of Aging (University of Chicago Press, Chicago) pp. 173–211.
Sialm, C. (2000), "Tax rate changes and the equity premium", Mimeo (Stanford University Department of Economics).
Siegel, L., and D. Montgomery (1995), "Stocks, bonds, and bills after taxes and inflation", Journal of Portfolio Management (Winter), 17–25.
Sims, T. (1995), "Taxation, optimization, and the January seasonal effect (and other essays in taxation and finance)", Ph.D. dissertation (Massachusetts Institute of Technology, Department of Economics).
Slemrod, J. (1982), "Stock transactions volume and the 1978 capital gains tax reduction", Public Finance Quarterly 10:3–16.
Slemrod, J., and T. Greimel (1999), "Did Steve Forbes scare the municipal bond market?", Journal of Public Economics 74:81–96.
Stephens, M., and J. Ward-Batts (2001), "The impact of separate taxation on the intra-household allocation of assets: evidence from the U.K.", Working Paper 8380 (National Bureau of Economic Research, Cambridge, MA).
Stiglitz, J.E. (1969), "The effects of income, wealth, and capital gains taxation on risk-taking", Quarterly Journal of Economics 83:262–283.
Stiglitz, J.E. (1983), "Some aspects of the taxation of capital gains", Journal of Public Economics 21:257–294.
Strickland, D. (1996), "Determinants of institutional ownership: implications for dividend clienteles", Mimeo (Ohio State University College of Business).
Summers, L.H., and V.P. Summers (1989), "When financial markets work too well: a cautious case for a securities transactions tax", Journal of Financial Services Research 3:261–286.
Talmor, E. (1985), "Personal tax considerations in portfolio construction: tilting the optimal portfolio selection", Quarterly Review of Economics and Business 25:55–71.
Tepper, I. (1981), "Taxation and corporate pension policy", Journal of Finance 36:1–13.
Tobin, J. (1978), "A proposal for international monetary reform", Eastern Economic Journal 4:153–159.
Trostel, P.A. (1993), "The effect of taxation on human capital", Journal of Political Economy 101:327–350.
Umlauf, S. (1993), "Transaction taxes and the behavior of the Swedish stock market", Journal of Financial Economics 33:227–240.
US Congressional Budget Office (1988), How Capital Gains Tax Rates Affect Revenues: The Historical Evidence (Congressional Budget Office, Washington).
Wolff, E.N. (1996), "Commentary", Tax Law Review 51:517–522.

Chapter 18

TAXATION AND SAVING*

B. DOUGLAS BERNHEIM

Stanford University, Stanford, CA and National Bureau of Economic Research, Cambridge, MA

Contents

* I am grateful to the National Science Foundation (Grant Number SBR95-11321) for financial support. During the course of this project, I have benefitted from discussions with numerous friends and colleagues, and my intellectual debts are too numerous to list. However, I am especially grateful to Alan Auerbach, William Gale, and Kenneth Judd. I am also indebted to Sita Nataraj for catching a number of errors, and for suggesting a variety of expositional improvements.

Handbook of Public Economics, Volume 3, Edited by A.J. Auerbach and M. Feldstein

Abstract

In this survey, I summarize and evaluate the extant literature concerning taxation and personal saving. I describe the theoretical models that economists have used to depict saving decisions, and I explore the positive and normative implications of these models. The central positive question is whether and to what extent specific public policies raise or lower the rate of saving. The central normative question is whether and to what extent it is desirable to tax the economic returns to saving. I also examine empirical evidence on the saving effects of various tax policies. This evidence includes econometric studies of the generic relation between saving and the after-tax rate of return, as well as analyses of responses to the economic incentives that are imbedded in tax-deferred retirement accounts. Finally, I also discuss several indirect channels through which tax policy may affect household saving by altering the behavior of third parties, such as employers.

Keywords

taxation, saving, personal saving, corporate saving, tax-deferred retirement accounts, pensions, interest elasticity of saving, optimal taxation, welfare costs of taxation

JEL classification: H20

1. Introduction

Recognizing the importance of saving as a determinant of both personal economic security and national economic performance, policymakers worldwide have become increasingly interested in developing effective strategies for stimulating (or in some cases discouraging) thrift. This interest has become particularly acute in the United States, where rates of saving are currently very low both by historical standards and in comparison to other developed countries. Concerns over low saving have led to a variety of policy proposals designed to stimulate thrift through the tax system, ranging from narrowly focused tax-deferred savings accounts to broad-based consumption taxation. Economic research has played an important role in the resulting public policy debates, and economists have weighed in on virtually all sides of the pertinent issues.

In this survey, I summarize and evaluate the extant literature concerning taxation and personal saving[1]. I describe the theoretical models that economists have used to depict saving decisions, and I explore the positive and normative implications of these models. The central positive question is whether and to what extent specific public policies raise or lower the rate of saving. The central normative question is whether and to what extent it is desirable to tax the economic returns to saving. I also examine empirical evidence on the saving effects of various tax policies. This evidence includes econometric studies of the generic relation between saving and the after-tax rate of return, as well as analyses of responses to the economic incentives that are imbedded in tax-deferred retirement accounts. Finally, I also discuss several indirect channels through which tax policy may affect household saving by altering the behavior of third parties, such as employers.

The remainder of the chapter is divided into five sections. Section 2 discusses theories of taxation and saving. It investigates the positive and normative implications of taxing the returns to saving under several variants of the life-cycle hypothesis, as well as under behavioral alternatives. Section 3 describes the available evidence on the generic relation between saving and the after-tax rate of return. It identifies two distinct approaches to measurement (estimation of consumption or saving equations, and estimation of consumption Euler equations), and it discusses the limitations of each. In Section 4, I examine evidence on the effects of opportunities to save through tax-deferred retirement accounts. This section focuses primarily on US tax policies, and includes detailed discussions of Individual Retirement Accounts (IRAs) and 401(k)s. Both IRAs and 401(k)s have accounted for large flows of saving, but there is heated controversy over the extent to which these flows represent new saving. In Section 5,

[1] National saving consists of two components: private saving and public saving. Private saving takes place among households (personal saving) and corporations (corporate saving). Public saving is the sum of budget surpluses (or deficits) for federal, state, and local governments. For the most part, this chapter concerns the impact of tax policy on the personal component of national saving. However, collateral effects on other components of national saving (e.g., changes in government revenue and shifts between corporate saving and private saving) are considered where relevant.

I shift attention to indirect links between taxation and household saving. I discuss the implications for household saving resulting from tax-induced changes in other aspects of the economic environment, including the size and scope of the pension system, the characteristics of employment-based pensions, the level of corporate saving, the availability of employment-based investment and retirement education, and the intensity with which financial institutions market and promote specific financial products. Section 6 concludes.

2. Theories of taxation and saving

For more than fifty years, the framework of intertemporal utility maximization has dominated economists' thinking about the tax treatment of saving. This framework traces its roots to Irving Fisher (1930), and lies at the heart of the Life Cycle Hypothesis (LCH) articulated by Modigliani and Brumberg (1954). Empirical tests of the LCH have yielded mixed results, leading some to modify the framework and others to reject it outright in favor of alternative approaches. In this section, I examine the positive and normative implications of the LCH, variants of the LCH, and alternative behavioral theories of tax policy and saving.

2.1. The life-cycle hypothesis

In the following discussion, I illustrate some pertinent implications of the LCH through a simple model. Imagine an individual who lives for a total of $T+1$ years, earning wages of w_τ in each year τ. This individual derives utility from consumption, c_τ, according to an intertemporally separable utility function of the form

$$\sum_{\tau=0}^{T} u_\tau(c_\tau)\rho^\tau, \tag{1}$$

where $\rho < 1$ represents a pure rate of time preference. The individual can alter the intertemporal allocation of resources by borrowing or lending. Let A_τ denote net asset holdings at the outset of period τ; for convenience, assume for the moment that $A_0 = 0$.[2] After receiving the wage w_τ and consuming c_τ, the individual is left with $A_\tau + w_\tau - c_\tau$. Prior to the start of period $\tau+1$, these investments earn pre-tax returns at the rate i. Capital-income taxes are applied symmetrically, so that interest received is taxed and

[2] This assumption is actually without loss of generality, since one can simply take the period-0 wage, w_0, to include the value of any initial assets.

interest paid is subsidized at the rate m [3]. Thus, for any given consumption path, asset holdings evolve as follows:

$$A_{\tau+1} = [A_\tau + w_\tau - c_\tau]\beta, \tag{2}$$

where

$$\beta \equiv 1 + i(1-m). \tag{3}$$

A consumption path is feasible as long as the individual dies with non-negative asset holdings [4]:

$$A_{T+1} \geqslant 0. \tag{4}$$

This restriction is equivalent to the requirement that

$$\sum_{\tau=0}^{T} c_\tau \beta^{-\tau} \leqslant W(\beta), \tag{5}$$

where

$$W(\beta) \equiv \sum_{\tau=0}^{T} w_\tau \beta^{-\tau} \tag{6}$$

represents the present discounted value of lifetime resources.

Behavior is governed by maximization of utility function (1) subject to restriction (5). It is useful for our current purposes to write optimal consumption as a function, $c_\tau(W,\beta)$, of the present discounted value of lifetime resources, W, and the discount factor, β. Using Equation (2), one can derive functions describing asset holdings, $A_\tau(W,\beta)$, along the optimal path. The associated level of saving, s_τ, is then given by the difference between total income (including investment returns) and consumption:

$$s_\tau(W(\beta),\beta) = \left(\frac{\beta-1}{\beta}\right) A_\tau(W(\beta),\beta) + w_\tau - c_\tau(W(\beta),\beta). \tag{7}$$

2.1.1. Positive analysis of taxation and saving

As is clear from Equations (3) and (7), conventional life-cycle models imply that changes in the capital-income tax rate, m, and in the pre-tax rate of return, i, both

[3] In practice, the tax system subsidizes interest payments to other parties by permitting individuals to deduct these payments from other income, subject to some limitations, prior to calculating taxes.

[4] In the special case where T is infinite, this inequality is replaced by the transversality condition.

influence saving by altering the after-tax rate of return, $i(1-m)$. The direction and magnitude of these effects are governed by the *interest elasticity of saving*.

In theory, the uncompensated interest elasticity of saving can be positive or negative, so saving can either rise or fall in response to an increase in the after-tax rate of return. This point is usually made in the context of a simple two-period model, where earnings are fixed and received entirely in the first period. In this setting, saving is equivalent to expenditure on second-period consumption. An increase in the after-tax rate of return amounts to an uncompensated reduction in the price of second-period consumption. The associated substitution effect shifts consumption towards the future (thereby increasing saving), while the associated income effect is usually assumed to increase consumption in both periods (thereby reducing saving). There is no theoretical presumption that either effect dominates. Indeed, with Cobb–Douglas utility (which implies fixed expenditure shares), a reduction in the rate of capital-income taxation has no effect on the level of saving, since the income and substitution effects offset exactly.

Further consideration of the two-period model suggests that the uncompensated interest elasticity of saving should depend on the distribution of earnings through time. In the standard Slutsky decomposition for the derivative of first-period consumption with respect to the price of second-period consumption, the income derivative is multiplied by the excess of second-period consumption over second-period earnings. Consequently, if second-period consumption exceeds second-period earnings, then the income effect associated with an increase in the interest rate results in greater first-period consumption. However, as one shifts earnings from the first period into the second period (holding the present discounted value of earnings constant so as not to alter consumption), the income effect shrinks, thereby enhancing the tendency for saving to rise in response to higher rates of return. When second-period earnings exceed second-period consumption, the household borrows in the first period; the income effect changes sign and reinforces the substitution effect.

These points remain valid even in more elaborate, multi-period life-cycle models, such as the one outlined above. Consider the effect on saving (equivalently, current consumption) of an unanticipated, permanent increase in the capital-income tax rate (m) at time $t=0$ [5]. Manipulation of the Slutsky equation allows us to decompose this into a substitution effect and an income effect:

$$\varepsilon^{u}_{0\beta} = \varepsilon^{c}_{0\beta} + \varepsilon_{0W}\left(\sum_{\tau=1}^{T}\tau\left(\frac{\beta^{-\tau}(c_\tau - w_\tau)}{W}\right)\right), \tag{8}$$

where $\varepsilon^{u}_{0\beta}$ is the uncompensated elasticity of period-0 consumption with respect to β, $\varepsilon^{c}_{0\beta}$ is the compensated elasticity of period-0 consumption with respect

[5] By focusing on period 0 and in assuming that the individual has no initial wealth (other than human capital), I am abstracting from possible wealth effects arising from asset revaluations.

to β, and ε_{0W} is the elasticity of first-period consumption with respect to lifetime resources (W)[6]. We know that $\varepsilon^c_{0\beta} < 0$, and normally $\varepsilon_{0W} > 0$. Focusing exclusively on the substitution effect, an increase in the after-tax rate of return (β) leads to a decline in consumption and an increase in saving. For earnings trajectories that give rise to no saving in any period ($c_\tau = w_\tau$ for all τ), the uncompensated interest elasticity of saving is governed entirely by the substitution effect; higher rates of return call forth more saving. As one shifts more resources towards the first period, initial saving becomes positive and subsequent saving becomes negative ($c_\tau > w_\tau$). The income effect counteracts the substitution effect, giving rise to smaller (potentially negative) interest elasticities of saving. As one shifts more resources away from the first period, initial saving becomes negative and subsequent saving becomes positive ($c_\tau < w_\tau$). In that case, the income effect reinforces the substitution effect, which suggests that households may reduce borrowing (increase net saving) sharply in response to an increase in the after-tax rate of interest.

To elucidate the relationship between the interest elasticity of saving and the structural parameters of the model, I will specialize to the class of utility functions that exhibit constant elasticity of intertemporal substitution:

$$u(c) = \frac{c^{1-\gamma}}{1-\gamma}. \tag{9}$$

Standard arguments imply that the optimal consumption profile satisfies the following Euler equation:

$$c_{\tau+1} = c_\tau(\beta\rho)^{1/\gamma}. \tag{10}$$

Equation (10) tells us that a change in the after-tax rate of return affects saving by altering the *slope* of the consumption trajectory. Moreover, the sensitivity of this response depends critically on $1/\gamma$, the intertemporal elasticity of substitution in consumption. In the extreme case of Leontief preferences ($1/\gamma = 0$), the slope of the consumption trajectory is entirely independent of β. Of course, this does not mean that the *level* of consumption is also independent of β. On the contrary, an increase in β reduces the present discounted value of any given consumption stream. If $W(\beta)$ is independent of β (which occurs if all earnings are received in period 0), a higher after-tax rate of return permits the individual to increase consumption in every period. With income fixed, this means that saving actually *declines* in response to a reduction in the rate of capital-income taxation. Thus, when $W = w_0$, a reduction in m can stimulate

[6] To derive this expression, note that $\partial c_0/\partial\beta = \sum_{\tau=1}^{T}(\partial C_0/\partial p_\tau)(\mathrm{d}p_\tau/\mathrm{d}\beta)$, where $C_0(p_1, \ldots, p_T, w_0, \ldots, w_T)$ describes optimal period-0 consumption as a function of the household's earnings stream and the implicit prices of consumption in later periods ($p_\tau \equiv \beta^{-\tau}$). Similarly, $\partial c_0/\partial\beta|_u = \sum_{\tau=1}^{T}\left(\partial C_0/\partial p_\tau|_u\right)(\mathrm{d}p_\tau/\mathrm{d}\beta)$. Note that the "substitution effect" is actually composed of many distinct substitution effects.

saving only if the slope of the consumption trajectory is sufficiently sensitive to β. This can only occur for higher values of $1/\gamma$.

Using Equations (5) and (10), one can obtain the following closed-form solution for initial consumption:

$$c_0(W,\beta) = \left(\frac{1-\lambda(\beta)}{1-\lambda(\beta)^{T+1}}\right) W, \tag{11}$$

where

$$\lambda(\beta) = \beta^{\frac{1-\gamma}{\gamma}} \rho^{\frac{1}{\gamma}}. \tag{12}$$

From Equation (11) it follows that, abstracting from the effect of β on the present discounted value of earnings (i.e. assuming that all earnings are received in period 0), $ds_0/d\beta$ has the same sign as $\lambda'(\beta)$. In the special case of Cobb–Douglas preferences (unitary elasticity of intertemporal substitution), $\gamma = 1$, so λ is independent of β, and saving is insensitive to the after-tax rate of return. For smaller elasticities of intertemporal substitution ($0 \leqslant 1/\gamma < 1$), $\lambda'(\beta) < 0$, so saving *falls* in response to an increase in the after-tax rate of return. Obviously, this includes the special case of Leontief preferences, discussed above. When the elasticity of intertemporal substitution exceeds unity ($1/\gamma > 1$), $\lambda'(\beta) > 0$, and saving *rises* in response to an increase in the after-tax rate of return. Thus, the sign of the pure price effect is indeterminate; there is no theoretical presumption that the interest elasticity of saving is positive. Moreover, with $W = w_0$, Cobb–Douglas preferences define the boundary between positive and negative elasticities.

When the household has positive future earnings ($W(\beta) > w_0$), Equation (11) implies that a change in β will also affect savings by altering the present discounted value of earnings. To study this effect in isolation, assume that $\gamma = 1$ (the Cobb–Douglas case), so that $\lambda'(\beta) = 0$ (the effect discussed in the previous paragraph vanishes). Provided that consumption is a normal good ($\partial c_0/\partial W > 0$), $dc_0/d\beta$ and $W'(\beta)$ have identical signs. Furthermore,

$$W'(\beta) = -\frac{W}{\beta} \sum_{\tau=1}^{T} \tau \left(\frac{w_\tau \beta^{-\tau}}{W}\right) \leqslant 0. \tag{13}$$

As long as the individual has some future earnings ($W > w_0$), the inequality in (13) is strict, which means that the interest elasticity of saving is necessarily *positive* in the Cobb–Douglas case. The intuition for this result is straightforward: an increase in the after-tax rate of return reduces the present discounted value of lifetime resources, thereby causing current consumption to fall, and current saving to rise.

As is clear from Equation (13), the size of $W'(\beta)$ depends on the timing of earnings. More specifically, the summation term is recognizable formally as the *duration* of the earnings stream $(w_0, w_1, \ldots, w_T)$. In words, duration is defined as a weighted average

of the times (τ) at which earnings are received, where the weights correspond to the fraction of total earnings (in present value) received at each point in time. When more earnings are received further in the future, duration is greater; the present discounted value of lifetime resources falls more in response to an increase in the after-tax rate of return, and so the associated increase in saving is larger.

If the duration of an individual's earnings stream is sufficiently large, then the interest elasticity of saving may be positive, even with Leontief preferences. Summers (1981) argues that $W'(\beta)$ is in fact quite substantial in realistically parameterized life-cycle models, and he suggests that this re-establishes a presumption in favor of the view that the interest elasticity of saving is positive and sizable. A number of authors have challenged this view. Evans (1983) demonstrates that the elasticities implied by these models are sensitive to the values of key parameters, including the assumed rate of time preference. Starrett (1988) exhibits sensitivity with respect to assumptions concerning the functional specification of utility[7]. As will be discussed in subsequent sections, it is also possible to overturn Summers' result by introducing liquidity constraints, uncertainty, and/or certain types of bequest motives.

Thus far, I have confined my remarks to tax policies that alter both the marginal and inframarginal returns to saving. It is also important to consider policies that do not have this feature. As will be discussed in Section 4, the US government has in the past attempted to stimulate saving through tax-deferred retirement accounts, which reduce the rate of taxation applicable to saving below some threshold level (the contribution limit). For the simple life-cycle model outlined in this section, saving within a tax-deferred account is a perfect substitute for other saving, and it also generates a higher return. Consequently, the model predicts that the contribution limit always binds. Even if desired saving is less than allowable contributions, individuals should reach the limit by borrowing or by shifting other assets. As a result, tax-deferred accounts do not alter the returns to saving on the margin. The reduction in the tax rate applicable to the returns from inframarginal saving amounts to a lump-sum windfall; the individual responds by increasing consumption and reducing saving.

2.1.2. Normative analysis of taxation and saving

Normative analyses of taxation and saving focus on two distinct but obviously interrelated sets of questions. First, should the government meet its revenue requirements in part by taxing the returns to saving? If so, how should it structure the tax, and what rates should it apply? Second, taking any particular tax system as a starting point, how

[7] In particular, one can reverse Summers' results by assuming that individuals have Stone–Geary utility functions of the form $u(c) = (c - \theta)^{1-\gamma}/(1-\gamma)$. Intuitively, some portion of saving is then motivated by the need to achieve a fixed target (the minimum consumption level θ) in every period. When the after-tax rate of return rises, the individual does not need to save as much to achieve this target. Consequently, when θ is large, the interest elasticity of saving tends to be small or negative.

large are the social gains or losses resulting from reforms that alter the tax burden on the returns to saving?

2.1.2.1. Optimal taxation of the returns to saving. The first set of questions concerns the role of capital-income taxation in an optimally designed tax system. The literature on optimal taxation contains a variety of pertinent results. For general background, see the related chapter 21 by Auerbach and Hines (2002) in this Handbook, or, for further discussion, the chapter by Chari and Kehoe (1999) in the Handbook of Macroeconomics.

There appears to be a presumption among many economists that capital-income taxes raise revenue less efficiently than taxes on consumption or wages. To understand the basis for this view, it is useful to start with the following simple model. Imagine an individual who lives for a total of $T+1$ years, and who derives utility from consumption, c_τ, according to the utility function $U(c_0,\ldots,c_T)$. For the moment, assume that the individual earns no wage income, but is endowed with initial assets A_0. Investments in period τ earn pre-tax returns at the rate i_τ between periods τ and $\tau+1$, and are taxed at the rate m_τ[8]. In addition, consumption is taxed at the time-invariant rate t[9]. The individual's budget constraint is then given by

$$\sum_{\tau=0}^{T}\left(\prod_{k=0}^{\tau-1}\frac{1}{1+i_k(1-m_k)}\right)c_\tau(1+t)\leqslant A_0. \tag{14}$$

If one sets m_τ equal to zero for all τ, Equation (14) simplifies to

$$\sum_{\tau=0}^{T}\left(\prod_{k=0}^{\tau-1}\frac{1}{1+i_k}\right)c_\tau\leqslant\frac{A_0}{1+t}. \tag{15}$$

It follows that, in this simple framework, a flat time-invariant consumption tax is equivalent to a non-distortionary lump-sum tax on endowments. In contrast, capital-income taxation is inefficient because it changes the relative prices of consumption in different periods, thereby rotating the budget constraint.

In the context of this same model, it is instructive to ask the following question. Suppose that consumption taxes are unavailable, so that the government must rely on distortionary capital-income taxes to raise revenue. How should it structure these taxes? While this question is somewhat artificial, it allows us to develop useful insights

[8] For simplicity, assume throughout that i_τ is fixed, so that the underlying production technology is linear.

[9] When time-varying consumption taxes and capital-income taxes are both available, there is some redundancy in the tax system. For example, one can replicate the effects of a time-invariant consumption tax with a system that taxes capital income at a constant rate, while taxing consumption at a decreasing rate over time.

concerning the optimal structure of capital-income taxes in more elaborate economic environments.

Setting $t = 0$ to reflect the absence of a consumption tax, we can rewrite the budget constraint as

$$c_0 + \sum_{\tau=1}^{T} q_\tau c_\tau (1 + \mu_\tau) \leqslant A_0, \tag{16}$$

where, in effect,

$$q_\tau \equiv \prod_{k=0}^{\tau-1} \frac{1}{1 + i_k} \tag{17}$$

is the producer price of consumption in period τ, and

$$\mu_\tau \equiv \prod_{k=0}^{\tau-1} \left(\frac{1 + i_k}{1 + i_k(1 - m_k)} \right) - 1 \tag{18}$$

is the effective tax rate on consumption in period τ.

When the model is reformulated in this way, it is immediately recognizable as a standard Ramsey optimal commodity-tax problem, where c_0 is the untaxed numeraire. One need only reinterpret standard results to characterize the optimal system of capital-income taxation. Under familiar (and commonly assumed) conditions, the government uses capital-income taxes temporarily, but then abandons them after some initial transition. More generally, it is possible to show that capital-income tax rates converge to zero with time (provided that T is sufficiently large)[10].

To understand these results, note that

$$\mu_\tau = (1 + \mu_{\tau-1}) \left(\frac{1 + i_{\tau-1}}{1 + i_{\tau-1}(1 - m_{\tau-1})} \right) - 1 \tag{19}$$

(where $\mu_0 \equiv 0$). Thus, a uniform commodity tax system ($\mu_\tau = \mu$, a constant, for all τ) is equivalent to a system in which capital income is taxed in period 0, but never thereafter ($m_0 > 0$, and $m_\tau = 0$ for all $\tau > 0$). A sufficient condition for the optimality of uniform

[10] The desirability of a zero long-run capital income tax rate emerges as a result in a variety of settings; see Diamond (1973), Auerbach (1979), Atkinson and Sandmo (1980), Judd (1985, 1999), Chamley (1986), Zhu (1992), Bull (1993), Jones, Manuelli and Rossi (1993, 1997) and Chari, Christiano and Kehoe (1994). Some of these papers analyze models with overlapping generations of (typically homogeneous) finite-lived agents, while others consider models with (sometimes heterogeneous) infinite-lived agents. The discussion in this section focuses on a simple case in which there is a representative agent whose horizon coincides with that of the economy, but it also includes some brief comments on the role of capital-income taxation in OLG models.

commodity taxation is that the utility function takes the form $U(x_0, \phi(x))$, where x_0 is the untaxed numeraire, x is the vector of taxed commodities, and ϕ: $\mathbb{R}^K \rightarrow \mathbb{R}$ (K being the number of taxed goods) is homothetic [Auerbach (1979)]. Note that these conditions are satisfied for the familiar (and commonly assumed) case of an intertemporally separable, isoelastic utility function [as in Equations (1) and (9)]. Consequently, with this formulation of utility, it is optimal to tax capital income once in the first period, but never again[11].

When preferences are *not* of the form described in the preceding paragraph, in general it will not be optimal to tax c_1 through c_T at a uniform rate. If the optimal commodity tax rate, μ_τ^*, rises with τ, then by Equation (19), it is optimal to tax capital income. Conversely, if the optimal commodity-tax rate falls with τ, then it is optimal to subsidize capital income.

Consider the infinite-horizon case where $T = \infty$. Imagine for the moment that the optimal capital-income tax rate is strictly positive and bounded away from zero in the long run. Then, by Equation (19), μ_τ^* converges to infinity for large τ. This implies that the distortion of future consumption rises without bound over time, which seems contrary to the usual principles of optimal taxation[12]. Recall in particular that the optimal commodity-tax rates are determined by compensated price elasticities. Provided that these elasticities converge to well-defined limits for large τ, one would expect μ_τ^* to converge to some finite limiting value, μ^*. This intuition is in fact correct.

[11] Technically, this solution is only valid when the present discounted value of the government's revenue requirement is not too large. When revenue requirements are substantial, the optimal value of μ may exceed i_0. According to Equation (19), this corresponds to an initial capital-income tax in excess of 100% ($m_0 > 1$). As long as individuals can invest in non-interest-bearing assets such as money, the *effective* tax rate on capital income can never exceed 100%, even if the statutory rate is greater. (If non-interest-bearing assets are nominal, then one can achieve an effective tax rate in excess of 100 percent, but there is still a maximum, and the analysis is qualitatively unchanged.) Thus, as we raise μ beyond i_0, a distortionary tax wedge appears between c_1 and other goods. Provided that preferences take the form $U(c_0, c_1, \phi(c_2, \ldots, c_T))$, where ϕ is homothetic, the new tax wedge will not disturb the conditions for optimality between c_2 through c_T, so uniform taxation of these goods will still be optimal. Thus, the solution would involve 100% capital-income taxation in the first period, positive capital-income taxation in the second, and no capital-income taxation thereafter. Of course, since $m_\tau \leqslant 1$ for all τ, Equation (19) also implies that there is a maximum effective commodity tax rate on consumption in every period τ: $(1+i_0)(1+i_1)\cdots(1+i_{\tau-1})-1$. If the revenue requirement is large enough, some of these other constraints will bind as well. However, the same logic applies: provided that preferences take the form $U(c_0, \ldots, c_\tau, \phi(c_{\tau+1}, \ldots, c_T))$ for all τ (where ϕ is always homothetic), constraints on effective tax rates for c_0 through c_τ will not disturb the optimality conditions that call for uniform taxation between $c_{\tau+1}$ through c_T. Notably, this condition is satisfied for the intertemporally separable, isoelastic case [Equations (1) and (9)]. Thus, the constrained solution always involves 100% capital-income taxation in periods 0 through some period $\tau-2$, positive capital-income taxation in period $\tau-1$, and no capital-income taxation thereafter. The use of capital-income taxation is therefore always transitory, and the period of transition is longer when revenue requirements are greater.

[12] Similarly, if the optimal capital-income tax rate is strictly negative and bounded away from zero in the long run, μ_τ^* converges to -1 for large τ, which also implies that the distortion of future consumption rises without bound.

It follows that the optimal rate of capital-income taxation may be positive or negative in the short run, but it converges to zero for the long run. Even if μ_τ^* converges to *several* limiting values, the associated capital-income taxes must still *average* zero in the long run [see Judd (1999) for further discussion].

The optimal tax problem described above is artificial in at least two respects: first, it assumes that tax instruments other than capital-income taxes are unavailable, and second, it assumes that taxes only distort decisions on the intertemporal margin. In practice, there are other taxes and other pertinent behavioral margins. Nevertheless, both the qualitative results and the associated intuition from the simple model are reasonably general.

To illustrate, modify the preceding model to incorporate a first-period labor–leisure choice, as well as a labor-income tax and a time-invariant consumption tax. Let L denote hours of leisure, $\bar{L}$ denote the individual's total endowment of time, w denote the hourly wage rate, and z denote the tax rate on labor income. For the moment, simplify by assuming that the individual has no initial assets ($A_0 = 0$). Then the budget constraint becomes

$$\sum_{\tau=0}^{T}\left(\prod_{k=0}^{\tau-1}\frac{1}{1+i_k(1-m_k)}\right)c_\tau(1+t) \leqslant (1-z)\,w(\bar{L}-L), \tag{20}$$

from which it is readily apparent that the consumption tax and the labor-income tax are equivalent. We can rewrite Equation (20) as

$$wL+\sum_{\tau=0}^{T} q_\tau c_\tau(1+\mu_\tau) \leqslant w\bar{L}, \tag{21}$$

where q_τ is, again, the producer price of consumption in period τ [see Equation (17)], and

$$\mu_\tau \equiv \left(\frac{1+t}{1-z}\right)\prod_{k=0}^{\tau-1}\left(\frac{1+i_k}{1+i_k(1-m_k)}\right)-1 \tag{22}$$

is the effective tax rate on consumption in period τ. This is again recognizable as a standard Ramsey optimal commodity-tax problem, where in this case L is the untaxed numeraire good and c_0 through c_T are the taxable goods. Provided that the utility function is of the form $U(L, \phi(c_0,\ldots,c_T))$ with ϕ homothetic, the optimal commodity-tax rates, μ_τ^*, are uniform, which requires a positive tax on either consumption or labor income ($z>0$ or $t>0$) and no taxes on capital income ($m_\tau = 0$ for all τ, including period 0). For models in which the individual potentially supplies labor in every period, a similar conclusion follows under an analogous condition[13]. Even if preferences do

[13] Assuming that the government can tax labor income at different rates in different years, capital income should not be taxed if utility function is of the form $U(L_0,\ldots,L_T,\phi(c_0,\ldots,c_T))$, where ϕ is homothetic.

not satisfy this condition, optimal capital-income taxes will still be zero in the long run (for $T=\infty$) provided that the μ_τ^* converge with τ to some limiting value, μ^* – a condition that holds with considerable generality.

The analysis becomes somewhat more complicated when the individual has initial assets ($A_0>0$). In that case, the budget constraint is

$$\sum_{\tau=0}^{T}\left(\prod_{k=0}^{\tau-1}\frac{1}{1+i_k(1-m_k)}\right)c_\tau(1+t)\leqslant(1-z)\,w(\bar{L}-L)+A_0. \tag{23}$$

Note that consumption taxation and labor-income taxation are no long equivalent. Indeed, by setting $t>0$ (*taxing* consumption), $z=-t$ (*subsidizing* labor), and $m_\tau=0$ for all τ, one can, in effect, create a non-distortionary tax on the initial endowment, A_0. This is usually regarded as an impractical solution since it ignores the incentive problems that arise if the government is unable to make a credible commitment *not* to expropriate accumulated capital. As a modeling strategy, it is therefore natural to assume that either the consumption tax or the wage tax is unavailable.

If the wage tax is unavailable, one can rewrite the budget constraint as

$$wL+\sum_{\tau=0}^{T}q_\tau c_\tau(1+\mu_\tau)\leqslant w\bar{L}+A_0, \tag{24}$$

where q_τ and μ_τ are defined as above [with $z=0$ in Equation (22)]. This is completely equivalent to the last case considered (with $A_0=0$). With the usual separability and homotheticity condition, taxation of capital income is undesirable. For $T=\infty$, optimal rates of capital-income taxation converge to zero in the long run provided that the optimal commodity-tax rate, μ_τ^*, converges to some limit.

It is natural to conjecture that the same results would hold when a labor-income tax is available and a consumption tax is not, but this is not quite correct. From Equation (23) it is evident that the labor-income tax combines a non-distortionary tax on a portion of the individual's endowment ($w\bar{L}$) with a distortionary leisure subsidy. Unlike a consumption tax, the labor-income tax does not extract revenue from the individual's financial endowment (A_0). In contrast, capital-income taxation provides the government with a mechanism for getting at the financial endowment, albeit at the cost of distorting decision-making on both the intertemporal margin and the labor–leisure margin. Clearly, capital-income taxation is unavoidable when $w\bar{L}$ is small relative to the government's revenue requirement and A_0. More generally, it is desirable to rely on both labor-income taxation and capital-income taxation, at least in the short run (even with the usual separability and homotheticity conditions), trading off the costs of the associated distortions against the benefits of tapping into different portions of the individual's endowment for the purpose of raising revenue.

We have already seen, however, that the government should not ordinarily rely on capital-income taxation in the long run even when its use is unavoidable in the

short run. Capital-income taxation necessarily entails distortions between current and future consumption, but one can avoid distorting consumption between different future periods by taxing investment returns only in transition. Similar principles apply in the case where a labor-income tax is available and a consumption tax is not. Provided that utility takes the form $U(L, c_0, \phi(c_1, \ldots, c_T))$ with ϕ homothetic, the government should use capital-income taxation to mimic a commodity tax that is uniform over $c_1, \ldots, c_T$ [14]. This is accomplished by taxing capital income in the initial period, and never thereafter [15]. Under relatively weak conditions, one can also guarantee more generally that, for economies with sufficiently long horizons, the optimal commodity-tax rates on c_T converge to a constant for large τ, which implies that the government should avoid capital-income taxation in the long run [Judd (1999)].

As is clear from this discussion, the avoidance of capital-income taxation in the long run has emerged as a major theme of the pertinent literature. It holds with considerable generality within a broad class of models. However, three qualifications are in order.

First, justifications for taxing or subsidizing capital income – even in the long run – may exist in more elaborate economic models. For example, Judd (1997) demonstrates that capital-income subsidies are optimal when firms exercise some degree of market power over intermediate capital goods (in effect, the subsidy offsets the private "tax"); conversely, capital-income taxes may be optimal in the long run when there is an untaxable source of pure profits that is related to the level of investment [Jones, Manuelli and Rossi (1993, 1997)]. As will be noted in Section 2.2.2, the existence of liquidity constraints may affect the desirability of capital-income taxation. Presumably, capital-income taxes or subsidies could also be optimal in the long run if the social benefits of investment activities (such as research and development) exceed the private benefits accruing to the investor.

Second, the optimal tax policy may not be time-consistent. Imagine, for example, that the government has access to taxes on labor income and capital income. Under appropriate conditions (see above), we know that the solution involves no capital-income taxation beyond the first period. Suppose, however, that the government re-optimizes each period. Provided that the individual holds positive assets, the re-optimized solution typically involves positive capital-income taxation in the short run. Consequently, the government is unwilling to follow through on its plan not to tax capital income after the initial period. In such situations, one can describe the government's choice as the equilibrium of a dynamic game. Under some conditions, it is still possible to construct equilibrium strategies that implement an efficient

[14] These restrictions are satisfied when, for example, one can write the utility function as $v(L) + u(c_0, \ldots, c_T)$, where $u(\cdot)$ takes the form described in Equations (1) and (9).

[15] Once again, depending on the magnitude of the revenue requirement, this may require an initial capital-income tax rate in excess of 100 percent, which is infeasible. In that case, there might be several transitional periods during which the government would tax capital income. See footnote 11 for further discussion.

tax scheme, but in other circumstances the rate of capital-income taxation may be either positive or negative, even in the long run [Benhabib and Rustichini (1997)].

Third, I have implicitly assumed throughout the preceding discussion that the representative household's planning horizon coincides with the horizon for the economy. Atkinson and Sandmo (1980), Auerbach (1979) and Diamond (1973) have studied the features of optimal tax systems in simple infinite-horizon models with overlapping generations of finite-lived individuals[16]. In these models, more restrictive conditions are required to guarantee that the optimal long-run tax on capital income is zero. Specific results depend on assumptions about the government's use of other policy instruments. When the government has sufficient control over the generational distribution of resources, the task of designing an optimal tax system is, in steady state, equivalent to the standard Ramsey tax problem for a representative finite-lived individual[17]. Though the limiting arguments mentioned in the preceding discussion no longer apply, the optimal capital-income tax rate is still zero in the long run if preferences are weakly separable into leisure and consumption, and homothetic in consumption.

When the government *cannot* optimize its use of debt, capital-income taxes play an important role in determining capital intensity. The steady-state welfare effects of capital-income taxation then depend on the divergence of initial steady-state capital intensity from the golden rule, and on the sensitivity of steady-state capital intensity to the after-tax rate of return. When capital accumulation is too low ($f'(k) > n$, where $f'(k)$ is the marginal product of capital and n is the population growth rate), the optimal tax structure reflects the benefits of setting capital-income taxes so as to encourage greater saving. Notably, in contrast to the standard Ramsey tax problem, the sign and magnitude of these benefits are governed by the *uncompensated* interest elasticity of saving, rather than the compensated elasticity. Since it is impossible to sign the uncompensated interest elasticity of saving as a matter of theory, the optimal

[16] Atkinson and Sandmo solve a problem wherein the government maximizes the discounted sum of individual lifetime welfares, and they examine the steady state of the solution. Auerbach solves a problem wherein the government maximizes steady-state welfare. The latter approach implies that the planner's social welfare function places no weight on the welfare of transitional generations. This favors policies that redistribute resources from transitional generations to steady-state generations, e.g., by moving the economy towards the golden-rule growth path. Such policies are not necessarily attractive when judged purely on the grounds of efficiency.

[17] For the most part, the literature considers models in which households live for two periods. In that setting, the equivalence result described in the text holds as long as the government can use public debt to achieve the desired steady-state capital stock. More generally, when households have T-period lifespans, the government needs $T-1$ redistributional instruments. The equivalence result also assumes that the government can implement age-specific taxes, for example by applying different tax rates to capital income earned by two distinct cohorts at the same point in time. The problem becomes more complicated when the government must apply the same tax rates to all cohorts at each point in time. However, the optimal long-run tax on capital income continues to be zero under the same conditions identified in the text.

tax structure can involve either a tax or a subsidy on capital income, even when there is too little capital accumulation in the initial steady state. In general, it is no longer desirable to refrain from taxing or subsidizing capital income even when the usual sufficient conditions (weak separability of preferences between consumption and leisure coupled with homotheticity in consumption) are satisfied.

2.1.2.2. The welfare costs of taxing the returns to saving. The second set of normative questions mentioned at the outset of Section 2.1.2 concerns the welfare effects of tax reforms. As a general matter, proposals to reform some arbitrary status quo by reducing or eliminating capital-income taxes can either raise or lower social welfare. Clearly, such proposals must inevitably reduce welfare when the status quo coincides with an optimal tax scheme involving positive taxes on capital income. Even when the optimal capital-income tax rate is zero, the welfare losses resulting from the taxation of investment returns can be either large or small, depending on the features of the economic environment.

Under certain conditions, one can approximate the welfare losses associated with the taxation of a consumption good by computing the area of a "Harberger triangle" [Harberger (1964)]. Since this area is proportional to the product of the square of the tax rate and the good's compensated demand elasticity, a small elasticity implies a small welfare loss. In the context of capital-income taxation, the pertinent behavioral margin involves the response of saving to a change in the after-tax rate of return. As will be discussed in Section 3, various studies have placed the uncompensated interest elasticity of saving at or near zero. If we take this evidence at face value and assume that the income effect in Equation (8) is small (e.g., because saving, s_0, is a small fraction of lifetime earnings, W), it is tempting to conclude that the compensated interest elasticity of saving must also be near zero, and to infer that the welfare costs of capital-income taxation are small. This inference is inappropriate: capital-income taxation may be highly inefficient even when compensated changes in the after-tax rate of return have little or no effect on saving [Feldstein (1978)].

To understand this point, imagine an individual who lives for two periods, supplying labor inelastically during the first period, and retiring prior to the second period. The relevant consumption goods are current consumption and future consumption, *not* current consumption and saving. Saving is related to *expenditure* on future consumption. To compute the size of the Harberger triangle, one must use the compensated elasticity of demand for second-period consumption with respect to the interest rate, rather than a compensated interest elasticity of saving[18]. Consequently, the Harberger triangle can be sizable even if the uncompensated interest elasticity of saving is zero and income effects are small.

[18] The notion of a compensated interest elasticity of saving is not even well-defined, since its size differs according to whether compensation is distributed in the first period or in the second period.

Feldstein (1978) uses the Harberger approximation to compute the welfare cost of capital-income taxation in a simple, two-period, representative-agent model. In the first period of life, the individual chooses labor supply and consumption; second-period consumption is determined as a residual. Assuming that the relevant *uncompensated* elasticities (the interest elasticity of saving, the labor supply elasticity, and the associated cross-price elasticities) are all zero, Feldstein finds that capital-income taxation entails substantial welfare losses. Specifically, when the initial tax rates on capital and labor income are both 40 percent, replacement of the capital-income tax with an equal-yield labor-income tax increases welfare by roughly 18 percent of tax revenue[19].

Notably, Feldstein's analysis abstracts from general-equilibrium effects, in that pre-tax factor returns (the wage rate and the interest rate) are taken as fixed. Other authors have explicitly considered the welfare costs of capital-income taxation in general-equilibrium growth models. This literature is divided into two segments: studies that employ models with infinite-lived households, and studies that employ models with overlapping generations of finite-lived households.

Chamley (1981) studies the welfare effects of replacing a capital-income tax with a non-distortionary lump sum tax in a model with a representative, infinite-lived household. He solves for the adjustment path from an initial steady state by linearizing the economy's equations of motion. Noting that the marginal deadweight loss of taxation is zero at the first-best allocation, he uses a quadratic approximation to compute the associated change in welfare. Under plausible parametric assumptions, he finds that, when labor supply is fixed, the welfare cost of capital-income taxation is approximately 11 percent of revenue when the tax rate is 30 percent, and 26 percent of revenue when the tax rate is 50 percent. The quadratic approximation implies that the welfare cost is roughly twice as high for the marginal dollar of revenue. These figures increase by as much as a third when Chamley allows for the possibility that capital-income taxes may also distort labor supply.

Judd (1987) studies a similar model, but improves upon Chamley's analysis in two ways: first, he considers experiments involving revenue-neutral changes in other distortionary taxes; second, he linearizes around steady states with positive taxes to obtain exact expressions for the marginal deadweight loss of taxation given any initial tax system. He finds that, on the margin, replacing capital-income taxation with labor-income taxation raises welfare for a broad range of estimated taste and technology parameters. Since the optimal long-run capital-income tax rate converges to zero for this class of models (see Section 2.1.2.1), this finding is understandable[20]. Judd's preferred calculations suggest that the welfare gain of an immediate and permanent

[19] It does not necessarily follow that the optimal income tax system involves no taxation of capital income under these parametric assumptions; Feldstein does not investigate this issue.

[20] For many of his parametric calculations, Judd also assumes that utility is additively separable in consumption and leisure, and that the consumption and leisure components are both homothetic. Chamley makes a similar assumption when he modifies his model to allow for variable labor supply. When

cut in capital taxation exceeds 25 cents on each dollar of revenue, and exceeds one dollar per dollar of revenue under plausible assumptions.

A somewhat different set of considerations governs the welfare effects of capital-income taxation in models with overlapping generations of finite-lived households. First, relative to models with infinite-lived agents, more restrictive conditions are required to guarantee that the optimal capital-income tax rate converges to zero in the long run (see Section 2.1.2.1). Consequently, there is less reason to believe *a priori* that a reduction in the capital-income tax rate will necessarily raise welfare. Second, unless the government adopts offsetting deficit policies, different tax systems may have different consequences for the intergenerational distribution of resources. It is important not to confuse these distributional effects with efficiency effects.

Taxes affect intergenerational distribution both directly and indirectly. Direct distributional effects result from different patterns of nominal incidence at fixed prices. For example, relative to a wage tax, a consumption tax distributes resources away from generations that are currently retired toward those that are currently working. Indirect distributional effects result from changes in equilibrium prices. For example, all else equal, an increase in capital accumulation during any period raises wages in subsequent periods by increasing the marginal productivity of labor, thereby benefitting later generations[21].

In overlapping-generations (OLG) models, tax policy affects capital accumulation (and hence intergenerational distribution) in two ways. First, saving may be sensitive to the after-tax rate of return. The associated effects of tax policy on capital accumulation and intergenerational distribution are governed by the uncompensated interest elasticity of saving. Second, there are general-equilibrium feedback effects from intergenerational distribution to capital accumulation (and hence back to intergenerational distribution). To illustrate, consider once again the choice between a wage tax and a consumption tax. Relative to a consumption tax, the wage tax leaves greater resources in the hands of current retirees, and less in the hands of current workers. Since retirees have higher marginal propensities to consume out of income, this tends to reduce

wage taxes are available but consumption taxes are not, these conditions imply that the optimal capital-income tax rate is exactly zero after some initial period of transition. Thus, when the government abandons the capital-income tax, it gives up an efficient levy on the initial capital stock, but this effect is swamped by the benefits of eliminating intertemporal distortions (at least for these parametric cases). Judd also considers parametric cases with non-separable utility for which optimal capital-income tax rates presumably converge to zero more gradually, and obtains similar results.

[21] A permanent increase in steady-state capital accumulation makes each steady-state generation better off only if the increase in labor productivity, and hence in after-tax wages, exceeds the required increase in saving. If the economy is on the "wrong" side of the golden-rule growth path (so that capital accumulation is inefficiently high, in the sense that $f'(k) < n$), then greater capital accumulation reduces steady-state welfare. In that case, tax policies that move the economy toward the golden-rule growth path can generate pure efficiency gains. However, it is generally believed that this is not the empirically relevant case. Movements toward the golden-rule growth path from below ($f'(k) > n$) raise issues of intergenerational distribution, rather than efficiency.

capital accumulation, thereby depressing wages in subsequent periods and distributing resources away from workers over a short-term horizon.

Diamond (1970) and Summers (1981) use OLG models to study the steady-state effects of capital-income taxation. Diamond obtains qualitative results for a model in which households live for two periods, while Summers attempts to quantify tax effects for a more realistic, parameterized model in which households live for fifty-five periods. The length of the household's horizon is important in this context because it affects the duration of the household's earnings stream, and hence the magnitude of the uncompensated interest elasticity of saving (see Section 2.1.1), which in turn governs the responsiveness of capital accumulation to changes in the after-tax rate of return. Summers' preferred calculations imply that steady-state welfare (expressed as a percentage of lifetime income) would rise by roughly 12 percent if capital-income taxes were replaced with consumption taxes, and by roughly 5 percent if capital-income taxes were replaced with wage taxes.

In evaluating Summers' results, one must bear several considerations in mind. First, he ignores the economy's transition path following tax reform[22]. The steady-state effects that he calculates are large because (i) the economy is below the golden-rule growth path ($f'(k) > n$), (ii) the model implies a substantial uncompensated interest elasticity of saving, and (iii) the effects of tax reform on capital accumulation are not offset by changes in government deficit policy. As discussed in Section 2.1.1, the implied value of the interest elasticity of saving – and hence the size of the associated steady-state welfare effect – is sensitive to parametric assumptions. More importantly, by focusing only on steady states, Summers' welfare calculations blend distributional effects with efficiency effects. It is important to remember that movements toward the golden-rule growth path benefit steady-state generations at the expense of transitional generations. If redistribution toward steady-state generations is desirable, the government could accomplish this objective in other ways (e.g., by running surpluses), without abandoning capital-income taxation.

Second, Summers assumes that households supply labor inelastically. Since this implies that the optimal capital-income tax rate is zero, reforms that eliminate capital-income taxes are inevitably welfare-improving. Consumption taxation and wage taxation are, in this model, equivalent to non-distortionary lump-sum taxation. Consequently, Summers does not examine policy experiments wherein the capital-income tax is replaced with another distortionary tax.

The equivalence of consumption taxes and wage taxes to lump-sum taxes, and hence to each other, may seem inconsistent with Summers' calculations, which imply that these two alternatives have very different steady-state effects. The explanation is that the switch from one system to the other would alter the timing of tax collection, but Summers does not permit offsetting changes in deficit policy. On average, consumption

[22] Summers (1981) cites an earlier unpublished version of his paper in which he examined the speed of transition, assuming myopic expectations.

occurs later in life than earnings. One can achieve a completely equivalent outcome, including an equivalent steady state, with a consumption tax or a wage tax levied at the same flat rate, provided that the government runs a higher debt with the consumption tax to compensate for the fact that it is collecting revenue later in the life of each individual. If one then eliminates this incremental debt (which Summers implicitly requires), steady-state capital accumulation will rise. Provided that the economy is initially below the golden-rule growth path, this increases steady-state welfare. The effect is, however, somewhat artificial, since the government could achieve the same outcome in the wage-tax setting by running a surplus.

Auerbach, Kotlikoff and Skinner (1983) [henceforth AKS; see also Auerbach and Kotlikoff (1987)] study a similar model, but improve upon Summers' analysis by allowing for variable labor supply, and by using computational methods to solve for the full dynamic path of the economy under rational expectations. Perhaps most importantly, they explicitly separate efficiency effects from distributional effects by examining two distinct types of tax-reform experiments. In the first type of experiment, tax rates are set to balance the government's budget period by period, and no government borrowing or lending takes place. This corresponds to Summers' approach. In the second type of experiment, tax rates are set to cover real exogenous government spending each period, but lump-sum transfers are used in combination with deficits and/or surpluses to alter the distribution of resources across generations. In particular, the authors hold fixed the utility of generations that are alive at the time of the tax reform, while distributing the net benefits or costs of the reform equally (expressed as percentages of lifetime income) across all subsequent generations. For the first type of experiment, results reflect a blend of distributional and efficiency effects, while the second type of experiment isolates efficiency effects.

Simulation results for the AKS model reveal a number of noteworthy patterns[23]. If the government were to replace the income tax with a consumption tax without adjusting deficit policy to fine-tune the intergenerational distribution of resources (a tax-reform experiment of the first type), the utility of the oldest initial cohorts would decline, but steady-state welfare (expressed as a percentage of lifetime income) would rise by roughly 6 percent. For a similar experiment involving the replacement of the income tax with a wage tax, the utility of the oldest initial cohorts would rise, but steady-state welfare would fall by roughly 4 percent. For tax-reform experiments of the second type, the welfare of all generations (other than those alive at the time of the reform) would rise by roughly 2 percent for a consumption tax, and fall by roughly 2 percent for a wage tax.

To understand the AKS results, it is helpful to begin with tax-reform experiments of the first type, for which steady-state results are directly comparable with Summers' analysis. Since labor supply is variable, the alternatives to capital-income taxation are

[23] These results presuppose an initial income tax rate of 30 percent. Auerbach and Kotlikoff (1987) provide results for a lower initial income tax rate (15 percent).

not distortion-free in the AKS setting, and consequently the steady-state outcomes with consumption taxation and wage taxation are considerably less attractive than in Summers' model. The steady-state ranking of consumption taxation and wage taxation continues to be driven by differences in the timing of revenue collection, coupled with the assumption that the government balances its budget period by period.

Differences between the transitional effects of consumption taxation and wage taxation originate from the divergent treatment of individuals who are already alive when the reforms are enacted. Since existing retirees earn no wages, they are plainly better off when the income tax is replaced with a wage tax, and worse off when it is replaced with a consumption tax. This differential treatment of the initial generations has two further effects. First, it implies that a consumption tax is less distortionary than a wage tax. In effect, the consumption tax supplements the wage tax with a non-distortionary capital levy. Since households must fund their consumption either from wages or from initial assets, the consumption tax base is strictly larger (in present-value terms) than the wage tax base, and the government can raise the same revenue with a lower tax rate. Like the wage tax, the consumption tax falls on labor and distorts the labor–leisure choice, but to a lesser extent since the rate is lower. Unlike the wage tax, the consumption tax also falls on initial assets, but this portion is non-distortionary since individuals cannot retroactively alter the labor earnings from which they accumulated their initial assets.

Second, the differential treatment of initial generations implies that consumption taxation promotes saving and capital accumulation in the short run, while wage taxation has the opposite effect. Within the life-cycle model, the marginal propensity to consume resources rises with age. Relative to an income tax, a consumption tax distributes resources away from older generations at the time of the reform, while a wage tax distributes resources towards these generations. The utilities of the oldest cohorts fall with consumption taxation, but younger generations benefit because higher capital accumulation raises wages and expands the tax base (permitting the government to apply even lower rates). In contrast, the utility of the oldest cohorts rises with wage taxation, but younger generations are adversely affected because lower capital accumulation depresses wages and contracts the tax base (requiring the government to impose even higher tax rates).

Now consider tax reform experiments of the second type, in which taxes are set to cover real exogenous government spending each period, but lump-sum transfers are used in combination with deficits and/or surpluses to alter the distribution of resources across generations. Under consumption taxation, transitional generations require compensation, so the steady-state outcome becomes less attractive (a welfare gain of 2 percent, instead of 6 percent). Under wage taxation, the government can extract compensation from the oldest cohorts, so the steady-state outcome becomes more attractive (a welfare loss of 2 percent, instead of 4 percent). The consumption-tax outcome is Pareto superior to the wage-tax outcome solely because the consumption tax incorporates a non-distortionary levy on existing capital, and thereby permits the government to impose a lower implicit tax rate on labor. Relative to

income taxation, consumption taxation has three effects: (i) it eliminates intertemporal distortions, (ii) it alters labor–leisure distortions, and (iii) it adds a non-distortionary levy on initial capital. The net impact of the first two effects is unclear[24], but the third effect is plainly beneficial. Relative to income taxation, wage taxation also has the first two effects, but it adds a lump-sum subsidy to initial capital. This third effect is plainly detrimental, since it requires the government to raise tax rates, aggravating the labor–leisure distortion. It is natural to wonder about the sign and magnitude of pure efficiency effects when one eliminates surprise capital levies and subsidies (the third effect) by considering fully anticipated tax reforms, but AKS do not undertake such experiments.

Subsequent research has refined, elaborated, and extended the work of Summers and AKS [see Auerbach and Kotlikoff (1987), Seidman (1983, 1984), Hubbard and Judd (1986), Starrett (1988), McGee (1989), Gravelle (1991a), Auerbach (1996) and Fullerton and Rogers (1993, 1996)]. Some of these studies are discussed in the next section, which considers variants of the life-cycle hypothesis.

2.2. Variants of the life-cycle hypothesis

Various studies have usefully extended the positive and normative analysis of capital-income taxation within life-cycle models to settings with additional realistic features. Chief among these features are bequest motives, liquidity constraints, and uncertainty. This section briefly summarizes these branches of the literature. For more detailed surveys, see Johnson, Diamond and Zodrow (1997) and Engen and Gale (1996a).

2.2.1. Bequest motives

Though there is widespread agreement that intergenerational transfers account for a significant fraction of household wealth, quantitative estimates vary widely. Kotlikoff and Summers (1981) conclude that roughly 50 to 80 percent of total wealth is due to intergenerational transfers, but subsequent studies tend to place this figure between 25 and 50 percent [see Aaron and Munnell (1992), Barthold and Ito (1992) and Gale and Scholz (1994a)]. To some extent, the dispute is definitional [see Modigliani (1988) and Kotlikoff (1988)].

[24] Although AKS do not solve for the optimal long-run capital-income tax rate when deficit policy is also optimized, there is no particular reason to believe that it is zero, since AKS depart from the assumptions that are known to generate this result. Specifically, AKS use a nested CES representation of preferences, with a parameter governing the substitutability between leisure and consumption within each period, and another parameter governing the substitutability between felicity in different periods. These preferences are not weakly separable in consumption and leisure. A natural conjecture is that the optimal capital-income tax rate is positive when contemporaneous consumption and leisure are substitutes (since this suggests that one should tax consumption more heavily during retirement), and negative when they are complements. However, without further analysis, it is impossible to know whether considerations arising from non-separability are quantitatively significant.

Theories of bequest motives fall into several distinct categories. One school of thought holds that bequests result from uncertainty concerning length of life coupled with restrictions on the availability of annuity insurance contracts [see Davies (1981)]. A second maintains that individuals care directly about the amount of wealth bequeathed to their heirs [see Blinder (1974), or Andreoni (1989)]. A third is predicated on the assumption that individuals have altruistic preferences, in the sense that they care directly about the utility or consumption of their heirs [see Barro (1974), or Becker (1974)]. A fourth depicts bequests as payments associated with transactions within families [see Bernheim, Shleifer and Summers (1985) or Kotlikoff and Spivak (1981)].

A number of studies have examined the empirical validity of these various alternatives. Collectively, the evidence points to a mixture of motives. Several authors have investigated the hypothesis that bequests are intentional, rather than accidental [Bernheim, Shleifer and Summers (1985), Hurd (1987, 1989), Bernheim (1991), Gale and Scholz (1994a)]. A number of studies have tested the altruism hypothesis by attempting to determine whether intergenerational transfers compensate for earnings differentials between generations and across children [Tomes (1981), Kurz (1984), Altonji, Hayashi and Kotlikoff (1992), and Laitner and Juster (1996)]. Bernheim and Bagwell (1988) argue that the altruism model leads inevitably to stronger, empirically untenable conclusions. Specific implications of exchange motives have also been examined [Bernheim, Shleifer and Summers (1985), and Cox (1987)]. All available theories have difficulty accounting for the robust empirical finding that more than two-thirds of US testators divide their estates exactly equally among their heirs [Menchik (1980), Wilhelm (1996)][25].

The implications of bequest motives for tax policy depend critically upon the type of motive that one assumes. For example, the taxation of bequests and inheritances is clearly non-distortionary if intergenerational transfers are accidental, but may have substantial efficiency costs if individuals have other motives. Different assumptions therefore lead to different implications concerning the desirability of including bequests in the consumption tax base, or inheritances in the wage tax base.

The interest elasticity of saving is also sensitive to one's assumptions about the nature of bequest motives. Standard formulations of the altruistic motive imply that the long-run interest elasticity of saving is much higher than in the absence of a bequest motive [Summers (1981), Evans (1983) and Lord and Rangazas (1992)]; indeed, the long-run partial-equilibrium interest elasticity of saving is infinite. In contrast, several studies have found that the interest elasticity of saving declines when one introduces accidental bequests [Engen (1994)] or preferences for bequests that are defined over the amount of wealth bequeathed rather than over heirs' consumption or utility [Evans

[25] Bernheim and Severinov (2000) argue that it is possible to account for the prevalence of equal division in a model with altruistic bequest motives if the division of bequests serves as a signal of the parent's relative affection for each child.

(1983), Starrett (1988), Fullerton and Rogers (1993)]. These are not general results, but depend instead upon the form of the utility function, and on the manner in which one recalibrates other parameters of the model when bequests motives are introduced[26]. In some instances, the interest elasticity of saving can even be negative. This might, for example, occur if an individual seeks to bequeath a fixed level of wealth: with a higher rate of return, less saving is required to reach the target.

Bequest motives also alter the welfare implications of capital-income taxation. If these motives are altruistic, then one can treat a sequence of finite-lived generations as a single, infinite-lived dynasty, and proceed as in Chamley (1981) and Judd (1987). If individuals' preferences are defined over the size of their bequest, the welfare effects of taxing the returns to saving become sensitive to the manner in which the model is calibrated. For such models, bequests are similar to consumption from the point of the testator, but they differ from consumption from the point of view of the economy because they add to capital accumulation. Consequently, when one incorporates bequests, one must recalibrate other parameters to replicate a baseline capital–labor ratio and interest rate. Evans (1983) recalibrates by adjusting the intertemporal elasticity of substitution, and finds that the introduction of a bequest motive significantly increases the impact of capital-income taxation on steady-state consumption. In contrast, Seidman (1984) recalibrates by adjusting the subjective discount rate, and finds that the welfare costs of capital-income taxation are relatively insensitive to the presence or absence of a bequest motive. Seidman also argues that the addition of a bequest motive reduces the transitional losses suffered by the initial generation of elderly individuals following a consumption-tax reform. This occurs for two reasons. First, when the model is recalibrated in the presence of a bequest motive, it implies less life-cycle saving, and hence less taxable consumption during retirement. Second, Seidman finds that, in the presence of a bequest motive, the elderly benefit from a slower rate of convergence to the new steady state.

2.2.2. Liquidity constraints

Up to this point, I have abstracted from liquidity constraints by assuming that individuals can borrow and lend at the same interest rate. The appropriateness of this assumption is debatable. There is a large empirical literature that attempts to assess the importance of liquidity constraints. One important branch examines data on asset holdings and the availability of credit, while another studies the sensitivity of consumption to income. A review of this literature is well beyond the scope of this chapter, but the interested reader can find citations, summaries, and evaluations in a variety of other places [see e.g., Attanasio (1995), Hubbard and Judd (1986), or Hayashi (1985)].

[26] The introduction of a bequest motive raises the steady-state capital–labor ratio and lowers the interest rate. If one adjusts other parameters (such as the intertemporal elasticity of substitution) to replicate baseline data, this will affect the interest elasticity of saving.

Liquidity constraints can in principle play an important role in determining the positive and normative effects of capital-income taxation. However, the nature of this role depends on one's assumptions concerning the characteristics of the market failure that gives rise to limitations on borrowing. The simplest approach is to model these limitations as exogenous non-negativity constraints on net worth (excluding human capital). One can justify this approach by appealing to transactions costs and/or the possibility of personal bankruptcy.

Since liquidity-constrained individuals do not alter their saving in response to small changes in the rate of return, the interest elasticity of aggregate saving tends to fall as binding credit constraints become more common. The introduction of an exogenous limitation on borrowing also implies that tax-deferred savings accounts can increase saving even in the presence of contribution limits. If desired saving is less than the limit and if the individual has no other wealth, then the limit must not bind. The availability of the tax-deferred account can therefore increase the individual's rate of return on the marginal dollar of saving – something that could not occur without liquidity constraints (see Section 2.1). Limitations on borrowing" also imply that saving in tax-deferred accounts may not be a perfect substitute for other saving (in contrast to the simple life-cycle model of Section 2.1). Since the government generally imposes significant penalties for early withdrawal, individuals sacrifice liquidity when they transfer assets into these accounts. If they anticipate a need to access savings prior to retirement (such as educational expenses for a child), they may prefer to save through other instruments[27]. It follows that individuals may choose to contribute less than the limit even when they have positive savings outside of tax-favored accounts, and that tax-favored saving may represent new saving even when contribution limits bind.

According to Hubbard and Judd (1986), the welfare costs of capital-income taxation in simulation models tend to be smaller (relative to the costs of labor-income taxation) in the presence of exogenous liquidity constraints. This reflects two considerations. First, since constrained individuals must deviate from their unconstrained optima, policies that exacerbate the severity and/or duration of the constraints are likely to have substantial, first-order efficiency costs. This effect is particularly pronounced when the intertemporal elasticity of substitution is low. A switch from capital-income taxation towards wage taxation reduces the consumption of constrained consumers, which produces a first-order reduction in welfare. Second, the potential efficiency gains from a reduction in capital-income taxation are smaller in the presence of borrowing limitations because constrained individuals do not alter their current consumption in response to a change in the after-tax rate of return. Due to these factors, the efficiency costs associated with an increase in the rate of labor-income taxation may exceed the efficiency gains resulting from a revenue-neutral reduction in the rate of capital-income taxation, even when the initial rate of capital-income taxation is

[27] If the anticipated needs are sufficiently far in the future, the individual may be better off saving through a tax-deferred account and paying the early withdrawal penalty.

substantial. It follows that the optimal capital-income tax rate may be positive [see also Aiyagari (1995)]. Similar issues arise with respect to consumption taxation. Though consumption tends to occur later in life than earnings, the two tax bases are identical during periods in which an individual encounters the borrowing constraint.

In some instances, it may be inappropriate to model liquidity constraints by introducing exogenous lower bounds on net worth. If credit-market failures result from informational asymmetries, the location of the constraint may be sensitive to other features of the economic environment. Under some conditions, changes in the timing of taxes over the life cycle can produce completely offsetting endogenous movements in borrowing constraints [see Hayashi (1985), as well as the discussions in Yotsuzuka (1987), and Bernheim (1987)]. In that case, a shift to wage taxation would not necessarily reduce current consumption.

2.2.3. Uncertainty and precautionary saving

Throughout the preceding discussion, I have assumed that households face no uncertainty with respect to their future incomes, exogenous expenses (such as medical costs), or any other factor. This is obviously a simplification. In practice, uncertainty plays a potentially important role in the life-cycle planning process, and gives rise to precautionary motives for saving. There is an extensive empirical literature that attempts to evaluate the importance of these motives. Various authors have examined the relationship between saving and measures of uncertainty, such as income variability and mortality risk. Others have relied on self-reported assessments of saving motives. A review of this literature is well beyond the scope of this chapter, but the interested reader is referred to the discussion in Engen and Gale (1996a).

The positive effects of capital-income taxation can change significantly when one introduces uncertainty. Unlike life-cycle saving, precautionary saving tends to be relatively insensitive to the after-tax rate of return. Consequently, when one adds uncertainty to a simulation model and recalibrates the model to replicate the same baseline capital–labor ratio and interest rate, the interest elasticity of saving can fall considerably. Using an overlapping-generations model similar to that of Summers (1981), Engen (1994) finds that, when earnings are stochastic, this elasticity is only one-tenth as large as it is when earnings are certain.

The introduction of uncertainty also has important implications concerning the positive effects of tax-deferred saving accounts. In the presence of credit constraints, uncertainty increases the value of liquidity, and thereby further reduces the degree of substitutability between liquid financial assets and illiquid tax-deferred saving. In the stochastic life-cycle model, the desire for liquidity is stronger among younger individuals, and the substitutability between tax-favored saving and other saving is lower. Consequently, as an individual ages, a shrinking fraction of tax-favored saving represents new saving. By the same token, contributions rise with age as the cost of illiquidity declines. Thus, the bulk of tax-favored saving is undertaken by individuals with a high degree of substitutability between tax-favored saving and other saving,

for whom a relatively small fraction of tax-favored saving represents new saving. Simulations suggest nevertheless that tax-favored saving accounts increase national saving significantly in the long run, but saving may decline in the short run as individuals fund their contributions from existing stores of wealth [Engen and Gale (1993, 1996a), Engen, Gale and Scholz (1994)].

The introduction of uncertainty also alters the normative effects of capital-income taxation. Using the overlapping-generations model mentioned above, Engen (1994) shows that steady-state welfare gains from replacing a capital-income tax with either a wage tax or a consumption tax are much smaller when income is stochastic. This finding reflects several factors. In Engen's model, the steady-state welfare cost of capital-income taxation is lower in the presence of uncertainty because the uncompensated interest elasticity of saving is smaller, and because it is necessary to recalibrate other parameters to compensate for the emergence of precautionary saving. Uncertainty also changes the welfare costs and benefits of labor-income taxation. Wage taxes mimic insurance by reducing the variability of after-tax income[28]. This beneficial effect is particularly pronounced when the labor-income tax is progressive. However, the associated reduction in uncertainty also mutes precautionary saving motives, thereby reducing capital accumulation and steady-state welfare.

Several authors have also explored normative aspects of capital-income taxation in stochastic models with infinite-lived agents. Given a particular realization of the state of nature, there is no reason to believe that it is optimal to tax consumption at an identical rate in any two consecutive periods. Consequently, the implied rate of capital-income taxation need not be zero, even in the long run. However, if the state of nature is not yet known, one might imagine that expectations about the optimal time-dated commodity-tax rates would converge to some limiting distribution over a long horizon. If these expectations are the same for periods t and $t+1$, then the sets of implied positive and negative capital-income tax realizations are mirror images of each other. It is therefore natural to conjecture that the optimal long-run *ex ante* capital-income tax rate is zero. Zhu (1992) shows that this conjecture is valid only under certain conditions, but Chari, Christiano and Kehoe (1994) find that the optimal long-run *ex ante* capital-income tax rate is approximately zero for plausible parameterizations of a stochastic simulation model.

2.3. *Behavioral theories*

In recent years, a number of economists have questioned the suitability of the life-cycle hypothesis for modeling the effects of tax policy on personal saving. Their concerns fall into two categories: issues related to bounded rationality, and issues related to self-control. I consider each of these in turn.

[28] Engen assumes that income variability is not insurable in the private sector, but he does not model the implied market failure explicitly. Depending upon the source of the market failure, the welfare gains from public insurance provision (e.g., through a wage tax) could be illusory.

Issues of bounded rationality arise from the complexity of intertemporal planning. To determine the solution of a standard life-cycle problem, an individual would require a high level of sophistication and extensive information on pertinent economic parameters. Yet much of the population appears ill-equipped to make even the most basic economic calculations [see Bernheim (1994a), or, for a general review of evidence on bounded rationality, Conlisk (1996)].

It is often argued that unsophisticated individuals may nevertheless act *as if* they solve complex mathematical problems. This view is particularly plausible when *either* (i) the activity in question is frequently repeated (so that the individual has the opportunity to experiment and learn), (ii) decisions taken by other individuals, as well as the consequences of these decisions, are both observable and pertinent (i.e. relevant vicarious experience is plentiful), or (iii) individuals recognize the need to obtain advice from qualified professionals, and have no difficulty obtaining this advice and monitoring its quality. Skeptics maintain that none of these conditions are satisfied in the context of the life-cycle planning problem. With respect to the first possibility, individuals usually retire only once – they have no opportunity to practice the life-cycle process. With respect to the second possibility, information on others' decisions is often poor. Moreover, since the consequences of these decisions are not fully known until well after an individual retires, and since 30-year-olds face very different economic conditions than the 90-year-olds whose consequences are fully known, vicarious observation of others tends to be either incomplete or of questionable relevance. Finally, with respect to the third possibility, unsophisticated individuals may be ill-equipped to evaluate the quality of information and advice provided by financial experts, or to evaluate experts' qualifications. In addition, they may not recognize or acknowledge the need for advice in the first place.

Formal models of bounded rationality typically proceed in one of several different directions [see Conlisk (1996) for a literature review]. Some impose structure on beliefs, for example by assuming a bias toward excessive optimism, a penchant for noticing salient or reassuring information, a tendency to forget information in the absence of rehearsal or corroboration, or a proclivity to update beliefs in a simplistic manner (e.g., through adaptive expectations). Others impose restrictions on decisions, limiting behavior to simple rules of thumb, such as saving a fixed fraction of income[29]. These restrictions are often empirically motivated. However, since they are not derived from generally applicable principles, this approach is necessarily somewhat *ad hoc,* and it fails to provide applied economists with a "tool kit" for addressing new problems. Other models envision costs to optimization [e.g., the notion of "satisficing", due to Simon (1955)]. While this approach appears to proceed from

[29] Notably, the advice of professional financial planners is often guided by extremely rough rules of thumb. According to the standard materials used for the curricula required to obtain the designation of Chartered Financial Consultant, "as a rule of thumb financial planners suggest that most families should plan to devote about 20 percent of their gross income to accumulation objectives" (Doyle and Johnson 1991).

general principles, the application of these principles is ultimately somewhat arbitrary. Instead of solving a particular optimization problem, one can certainly formulate and solve an alternative meta-problem that incorporates costs of computation. However, it is no less objectionable to assume that an individual can costlessly solve this meta-problem, than to assume that the individual can costlessly solve the original problem. Any coherent treatment of computational costs would therefore appear to lead to an infinite regress [Lipman (1991)].

The second issue – self-control – refers to the ability to follow through on intertemporal plans that require an individual to forego short-term gratification. While the life-cycle hypothesis implicitly assumes that self-control is perfect, a large body of psychological research suggests that imperfect self-control lies at the heart of many intertemporal decision-making problems [see e.g., Ainslie (1975, 1982, 1984, 1992), Maital (1986), Furnham and Lewis (1986), Schelling (1984), Thaler and Shefrin (1981), Shefrin and Thaler (1988) and Hoch and Lowenstein (1991)].

One can formalize problems of self-control in a number of different ways. Thaler and Shefrin (1981) propose a model in which an individual decision-maker consists of two distinct "selves" – a farsighted, patient "planner" and a shortsighted, impatient "doer". The planner can keep the doer in check only by expending costly effort ("willpower"). Laibson (1994a,b, 1996) analyzes a class of models in which problems with self-control arise directly from time-inconsistent preferences[30]. In contrast to the LCH, Laibson's formulation of the intertemporal planning problem assumes that an individual becomes less willing to defer gratification from period t to some period $s > t$ once period t actually arrives. As a result, the individual is typically unwilling to follow through on an optimal intertemporal plan. One can derive Laibson's model from a multiple-self framework similar to that of Thaler–Shefrin by assuming that the "planner" and the "doer" strike an efficient bargain in every period.

Existing models of self-control have at least one serious drawback: their solutions are significantly more complex than those of standard life-cycle problems. For example, the application of Laibson's framework requires one to solve for the equilibrium of a dynamic game played between an individual's current "self" and all of his or her future incarnations. Thus, in "solving" the problem of self-control, these frameworks accentuate the problems associated with cognitive limitations.

2.3.1. Positive analysis of taxation and saving

One can find a fair number of references to alternative behavioral hypotheses in otherwise conventional analyses of tax policy [see e.g., the discussions of IRA advertising in Venti and Wise (1992), and of "false" contribution limits in Feenberg

[30] Laibson's approach is motivated by psychological research, indicating that rates of time preference are approximately "hyperbolic" [see e.g., Ainslie (1992)]. His analysis of behavior is an outgrowth of earlier work on time-inconsistent preferences by Strotz (1955), Phelps and Pollak (1968) and others.

and Skinner (1989)]. Yet these references are usually haphazard, and mentioned in a rather *ad hoc* way as possible explanations for otherwise puzzling phenomena. With rare exceptions, alternative behavioral hypothesis have not been used as frameworks for organizing lines of inquiry concerning the effects of taxes on saving[31].

Certain behavioral hypotheses have clear implications concerning the effects of tax policy on saving. Consider, for example, the possibility that advice from professional financial advisors has a significant impact on behavior. If this view is correct, then to say something about the interest elasticity of saving, one should examine the nature of advice and determine how this advice changes with a change in the after-tax rate of return. The most common retirement-planning technique involves setting some fixed target for retirement (usually derived from an arbitrary earnings replacement rate) and computing the annual inflation-adjusted contribution to savings sufficient to achieve this target [see Doyle and Johnson (1991)]. The resulting interest elasticity is negative because higher rates of return make it easier to accumulate the resources required to reach the target.

While the implications of other behavioral hypotheses are often less clear, some alternatives lend themselves to formal analysis. Laibson, Repetto and Tobacman (1998) examine the steady-state effects of providing consumers with opportunities to save through accounts that resemble 401(k)s (contributions are deductible, earnings accumulate tax-free, and early withdrawals are penalized). Their model is similar to that of Engen, Gale and Scholz (1994), except that the specification of consumer preferences allows for hyperbolic discounting. According to their simulations, the steady-state rate of national saving rises significantly in the presence of tax-deferred retirement accounts, and the effect is roughly 30 percent larger when consumers have hyperbolic preferences (relative to the baseline case in which consumers have standard exponential preferences).

Though the literature on behavioral alternatives to the LCH contains few sharp predictions concerning the positive effects of tax policy on saving, it does suggest a number of pertinent qualitative principles. Specifically, taxes can change perceptions concerning the costs and benefits of saving, they can affect the feasibility of self-control by influencing the structure of private behavioral rules, and they can have an impact on personal saving indirectly by altering the decisions of third parties. I will elaborate on each of these possibilities in turn.

(i) *Perceptions of the costs and benefits from saving.* When saving incentives are in place, boundedly rational individuals may be more likely to learn that others regard the benefits of saving as important. For example, the availability of a 401(k) may stimulate conversations about contributions and investments, and thereby produce "peer-group"

[31] Exceptions include Thaler (1994), Bernheim (1994a, 1995, 1997a), Laibson (1996, 1998) and Laibson, Repetto and Tobacman (1998).

influences involving both demonstration and competition[32]. Likewise, tax incentives may stimulate promotional and educational activities that underscore the long-term benefits of saving (see the discussion of third-party activities later in this section, as well as Sections 5.4 and 5.5). The very existence of a pro-saving policy may indicate that "authorities" perceive the need for greater thrift. Likewise, individuals may attach significance to contribution limits (expressed either as fixed amounts or as fractions of compensation), on the grounds that these limits reflect the judgement of experts.

By segmenting retirement saving from other forms of saving, certain kinds of tax-favored accounts may make it easier to monitor progress towards long-term objectives. Information on total accumulated balances is usually provided automatically, or is readily available. Thus, individuals have a convenient yardstick for measuring the adequacy or inadequacy of their thrift. For those who save little, this may have the effect of making the costs of short-sightedness more explicit.

Thaler and Shefrin's behavioral life-cycle model assumes that the planner values saving, while the doer does not. In this setting, one imagines that tax incentives might affect saving by altering the *planner's* perceptions of costs and benefits. However, it is also possible that saving incentives might affect behavior by influencing the *doer's* perceptions. Scitovsky (1976) has raised the possibility that some individuals may view saving as a virtuous activity in and of itself, without any explicit contemplation of future consequences (see also Katona 1975). Pro-saving policies may promote this outlook by reinforcing the notion that, as something worthy of encouragement, saving is intrinsically rewarding and immediately gratifying.

Under certain circumstances, contributions to tax-favored accounts may also instill the perception that saving yields more concrete short-run benefits. By making tax-deductible contributions to a tax-favored account (when permitted), an individual can reduce the amount of taxes owed in the current year, or increase the size of his or her refund. Feenberg and Skinner (1989) have argued that the prospect of writing a larger check to the bank and a smaller check to the IRS may be particularly appealing on psychological grounds. Since the basis of this appeal (beating the IRS today) is a form of instant gratification, up-front deductibility may weaken the doer's opposition to thrift. This observation has potentially important implications concerning the choice between "front-loaded" and "back-loaded" plans. In a front-loaded plan, contributions are deductible and withdrawals are fully taxable; in a back-loaded plan, contributions are not deductible and withdrawals of *principal* are not taxable. The preceding discussion suggests that front-loaded plans may be more effective, since they may coopt impatient selves with the immediate reward of a current-year tax deduction. In contrast, under the LCH, individuals should prefer front-loaded plans to back-loaded plans if and only if they expect their marginal tax rates to fall.

[32] There is considerable evidence that economic decisions in general are strongly affected by peer-group effects. See e.g., Whyte (1943), Rainwater (1970), Stack (1974), or Jones (1984). For evidence on peer effects in the context of 401(k) plans, see Duflo and Saez (2000).

(ii) *Private rules.* The literature on self-control emphasizes the use of "private rules". Hoch and Lowenstein (1991) argue that individuals overcome impulsive inclinations by attaching global significance to small transgressions of these rules. For example, individuals may stake some aspect of their personal self-worth on their ability to follow a self-imposed rule; the benefits of breaking the rule in any isolated instance are counterbalanced by the loss of self-esteem. Similarly, an individual may construe transgressions of a rule as evidence that he or she will never be able to follow similar rules; consequently, the short-term gains from deviation are weighed against the losses associated with all related failures of self-discipline, now and in the future. With hyperbolic discounting, behavior of this kind is sustainable as an equilibrium of the intertemporal game played between an individual and his or her future incarnations [Laibson (1994a)].

Saving incentives may facilitate the formation of effective private rules in three ways. First, they may provide a natural context for developing rules concerning the level of saving. Possible rules could include always "maxing out" on tax-favored contributions, or always contributing some smaller amount to tax-deferred plans. Certain plans, such as 401(k)s, actually provide participants with limited ability to commit themselves to these rules for short periods of time.

Second, individuals may also develop private rules regarding the allowable uses of funds that they have previously placed in tax-favored accounts. For example, they might promise themselves that they will not withdraw these funds for any purpose short of a dire emergency. This phenomenon relates to the notion of "mental accounting" discussed by Shefrin and Thaler (1988). The existence of penalties for early withdrawal may help the individual establish and enforce barriers around tax-favored accounts. Somewhat paradoxically, these barriers may be high precisely because impatient selves (doers) have a strong aversion to paying immediate penalties. Anticipating a possible future loss of self-control, an individual may actually be more likely to contribute to a tax-favored account that provides a credible mechanism for precommitment. In contrast, under the life-cycle hypothesis, restrictions on early withdrawals reduce the likelihood that individuals will be willing to make contributions.

Third, as mentioned above, tax-favored savings accounts may make it easier to monitor progress toward long-term objectives. Effective monitoring is essential for the enforcement of private rules. According to Thaler and Shefrin (1981), "simply keeping track seems to act as a tax on any behavior which the planner views as deviant".

(iii) *Third-party activities.* Non-neutralities in the tax system may stimulate activities by "third parties" – that is, parties other than the individuals who benefit directly from the tax provisions, such as employers or vendors of tax-favored investments products. These activities may in turn affect the level of personal saving through either life-cycle or psychological channels.

The most obvious example of this phenomenon is the private pension system. As will be discussed in Section 5.1, the tax benefits accorded pensions probably account, at least in part, for their popularity. When an employer offers a traditional defined-benefit or defined-contribution pension plan, saving automatically increases unless

the individual takes steps to negate this effect. Pure life-cycle decision-makers would pierce the "pension veil" and treat the accrued value of pension benefits as a close substitute for other long-term saving. Even so, mandatory pensions may increase the saving of some households by forcing them to undertake more long-term saving than they would otherwise choose. Contributions to pension plans may also represent incremental private saving under various alternative behavioral hypotheses. Households may pierce the pension veil imperfectly, they may track pension accruals in different "mental accounts" than other long-term saving, or the mere presence of a pension plan may make them more aware of retirement issues.

Selective saving incentives may also have subtle effects on the features of pension plans. For example, 401(k) plans have historically received favorable tax treatment only if they satisfied non-discrimination requirements regarding the relative levels of benefits provided to highly compensated and non-highly-compensated employees. Rather than risk losing tax-favored status, many firms have taken steps to increase the participation and contributions of non-highly-compensated employees, and/or to decrease the contributions of highly compensated employees [Garrett (1995)]. These steps often included provisions whereby firms matched employee contributions, and the adoption of retirement education programs. These kinds of plan features have the potential to affect overall saving by eligible workers. Education may be particularly effective if low saving results from a failure to appreciate financial vulnerabilities.

Selective incentives may also encourage the vendors of tax-favored savings vehicles to advertise and promote their products actively. These promotional efforts may serve an educational function, or simply focus public attention on retirement income security. For example, the expansion of eligibility for IRAs to all taxpayers in 1981 was accompanied by a great deal of advertising and media fanfare.

The distinctive positive implications of the behavioral framework are perhaps most apparent when one considers the choice between broad-based policies for promoting saving, such as consumption taxation, and more limited strategies, such as IRAs. IRAs and other narrowly focused programs raise the *marginal* after-tax rate of return only for particular types of saving, and only if this saving does not exceed contribution limits. In contrast, a shift to broad-based consumption taxation would raise the marginal after-tax rate of return for all households, irrespective of the amount saved or the reason for saving. Provided that the interest elasticity of saving is positive, the LCH therefore leads us to expect that saving would increase more in response to consumption taxation than to narrower programs. Yet some of the behavioral considerations discussed in this section suggest the opposite. Narrow measures can focus attention on a single issue (such as the adequacy of saving for retirement), expose individuals to information concerning the importance of saving, provide a natural context for the development and enforcement of private rules, and promote the growth of pro-saving institutions. Contribution limits in particular may actually stimulate saving if they validate specific targets, provide natural focal points for the formation of private rules, or make it easier to monitor compliance with these rules. Conversely, a broad-based consumption tax could undermine the narrow focus on specific objectives that may be essential for the

exercise of self-control. It would remove one of the primary reasons for compensating workers through pension plans, and it would eliminate the special feature of particular financial instruments (such as IRAs and life-insurance policies) that make them especially marketable. It would also eliminate the quirky aspects of the tax system that subtly promote activities such as employee retirement education.

Before moving to a discussion of the evidence on taxation and saving, it is also important to emphasize that, depending upon whether one adopts the perspective of the LCH or some behavioral alternative, one may be inclined to draw very different positive inferences from the same set of empirical findings. As an example, consider the generalizability of evidence on the interest elasticity of saving. Within the context of the LCH, all saving incentives motivate changes in behavior through the same fundamental mechanism: an increase in the after-tax rate of return alters the intertemporal terms of trade. Measurement of a "generic" interest elasticity of saving therefore emerges as a central research priority. Alternative behavioral hypotheses allow for the possibility that the interest elasticity of saving may vary according to context, and that households may respond (both positively and negatively) to aspects of tax-incentive programs that are not directly related to the after-tax rate of return. In that case, measurement of the interest elasticity of saving in one context may shed little light on the effectiveness of tax policy in another context.

2.3.2. Normative analysis of taxation and saving

Proponents of pro-saving policies frequently argue that the prevailing rate of saving is "too low", and that individuals are providing inadequately for their futures [see e.g., Bernheim (1997b)]. Although it is possible to make sense of these claims within the context of the LCH, further clarification is required. A deliberate, forward-looking life-cycle planner carefully weighs the costs and benefits of saving. While impatient individuals may appear to save too little from the perspective of those with greater patience, this is merely a reflection of preferences. A traditional guiding principle of US economic policy is respect for free choice and diversity of tastes. A devotee of classical music might similarly deplore popular musical genres, but this is hardly an argument for subsidizing recordings of Stravinsky.

Once one steps away from the LCH, it is much easier to make sense of the claim that individuals save too little (e.g., if profligacy results from a failure to understand financial vulnerabilities, or from an unintended break-down of self-control). Moreover, the welfare gains associated with these policies are likely to be much larger than those implied by the LCH. In general, variations in consumption have greater effects on welfare when initial choices are farther removed from an optimum. Thus, under the LCH, the welfare costs of a small tax on capital income are second-order, and the welfare costs of a larger tax are limited by the extent to which that tax induces a departure from the optimum. In contrast, under alternative behavioral hypotheses, an individual may depart substantially from his or her optimum even in the absence of a tax. Thus, the marginal benefits from stimulating saving are potentially much greater.

According to Laibson's (1996) simulations, customers with hyperbolic preferences are willing to sacrifice nine-tenths of a year's worth of income to induce the government to implement optimal revenue-neutral saving incentives.

3. Evidence on responses to changes in the after-tax rate of return

Much of the literature on the relation between taxation and personal saving attempts to measure the interest elasticity of saving without reference to a specific policy. Studies of this kind implicitly assume there is a well-defined *generic* interest elasticity of saving. While this premise is valid under the LCH, some behavioral alternatives suggest that it is impossible to separate behavior meaningfully from institutional context (see Section 2.3).

The magnitude of the interest elasticity of saving is inherently an empirical issue; as discussed in Section 2, even the sign of this elasticity is theoretically ambiguous. The extant literature reflects two distinct approaches to measurement. One approach involves the estimation of functions describing either consumption or saving. The second approach involves the recovery of structural preference parameters through the estimation of consumption Euler equations. I will discuss each of these in turn. The interested reader may also wish to consult Elmendorf's (1996) survey for additional details.

3.1. The consumption/saving function approach

The earliest approach to measuring the interest elasticity of saving involved the estimation of a consumption function or saving function featuring an interest rate among the list of explanatory variables. Since the initial work of Wright (1969), this approach has yielded a variety of elasticity estimates, ranging from essentially zero [Blinder (1975), Howrey and Hymans (1978), Skinner and Feenberg (1989)] to 0.4 [Boskin (1978), Boskin and Lau (1978)]. This range is somewhat misleading, since the estimates tend to cluster near zero. There has been considerable discussion in the literature concerning the sources of the discrepancies between these various estimates [see e.g., Sandmo (1985)], with particular attention being given to the proper measurement of the real after-tax rate of return[33].

This approach has been criticized on the grounds that explanatory variables such as disposable income and the interest rate are potentially endogenous. A more fundamental problem follows from the "Lucas critique" of reduced-form empirical models [Lucas (1976)]. Since the relation between consumption (saving) and interest rates depends on expectations (which in turn result from the broader economic

[33] Given the complexity of the US tax system, some have even questioned whether it is possible to summarize the intertemporal terms of trade with a single number [Balcer and Judd (1987)].

context), there may not exist anything that one could properly regard as a stable saving or consumption function. If changes in the interest rate are correlated with changes in expectations about future resources or economic conditions, this will confound efforts to identify the interest elasticity of saving. In short, the historical relation between saving and the after-tax rate of return may provide a poor basis for forecasting the manner in which saving might respond to future changes in tax policy.

An inspection of historical US data reinforces this concern. Low-elasticity estimates are largely attributable to data from the 1970s, during which saving was relatively high and *ex post* real rates of return were very low. Since the 1970s were unusual in many other respects that might have affected expectations, this limited experience provides a questionable basis for forecasting future changes in saving. Unfortunately, the historical record does not offer a "clean" macroeconomic experiment involving a change in the rate of return and no change (or, at least, a known change) in expectations, from which one might *directly* infer the interest elasticity of saving.

3.2. The Euler-equation approach

As discussed in Section 2, a variety of studies compute interest elasticities of saving, as well as the welfare costs of alternative tax systems, in hypothetical economies populated by optimizing agents. Various authors have used these models to map estimates of structural preference parameters into estimates of elasticities and efficiency effects.

Naturally, the positive and normative effects of capital-income taxation depend upon a large number of economic parameters. As noted in Section 2.1.1, one critical preference parameter is the intertemporal elasticity of substitution in consumption [$1/\gamma$ in Equation (9)]. Equation (10) (the consumption Euler equation) implies that this parameter governs the rate at which the slope of the consumption trajectory responds to changes in the after-tax rate of return.

Note that one can rewrite Equation (10) as follows:

$$\frac{\Delta c_\tau}{c_\tau} \approx \frac{1}{\gamma} r - \frac{1}{\gamma}\delta, \tag{25}$$

where r is the real after-tax rate of return ($i(1-m)$ in the model of Section 2.1.1), and $\rho \equiv 1/(1+\delta)$[34]. Equation (25) suggests that it is possible to estimate the intertemporal elasticity of substitution by regressing the fractional change in consumption on the real after-tax rate of return. As a formal matter, since we derived Equation (25) from a model with no uncertainty, it is a deterministic relation and not a stochastic

[34] To obtain this expression, take logs of both sides of Equation (10) and use $\ln(1+x) \approx x$.

regression equation. Under some conditions [see, e.g., Hall (1988), or Deaton (1992)], Equation (25) generalizes in the presence of uncertainty to the following expression:

$$\frac{\Delta c_\tau}{c_\tau} \approx \frac{1}{\gamma} r - \frac{1}{\gamma}\delta + \mu + \varepsilon, \tag{26}$$

where μ depends on the variance of errors in forecasting consumption growth and ε is a random disturbance.

In principle, it is possible to estimate Equation (26) and recover $1/\gamma$ from the coefficient of the real after-tax rate of return. The contemporaneous value of r may be correlated with the error term, either because it is determined endogenously with the change in consumption or because it is associated with new information that affects the level of consumption. However, theory implies that ε is orthogonal to all information available prior to time τ, including past disturbances. Lagged variables are therefore necessarily exogenous, and make ideal instruments for r.

The procedure described in the preceding paragraph finesses a number of problems that arise with respect to the estimation of functions explaining aggregate consumption and/or saving. It provides a theoretically coherent treatment of endogeneity issues, it identifies structural preference parameters, and it avoids the estimation of reduced-form coefficients that are confounded by expectational and informational effects. Naturally, some problems remain (e.g., difficulties associated with the measurement of an after-tax real rate of return), and a number of new problems emerge (see below).

The interpretation of the coefficient of r in Equation (26) as the intertemporal elasticity of substitution is, of course, model-specific. As Hall (1988) notes, the standard life-cycle model makes an automatic connection between this intertemporal elasticity and the coefficient of risk aversion, whereas no connection appears to exist in practice. Although Hall exhibits one specification of utility that breaks this connection while still generating an Euler equation with an identical interpretation, there is no guarantee that this conclusion follows for other specifications. Other models obscure the structural interpretation of Equation (26), thereby rendering the approach vulnerable to the Lucas critique. For example, in the presence of uncertainty and liquidity constraints, individuals may engage in "buffer stock" saving. The expected desirability of next period's consumption – and hence the slope of the optimal consumption profile – may then depend on all factors affecting the probability that the individual will run out of liquid wealth, including expectations about future income. If one moves to other behavioral hypotheses, estimates of Equation (26) may have no structural underpinnings.

For the United States, there has been relatively little historical correlation between the growth rate of aggregate consumption and measures of the after-tax rate of return. Consequently, estimates of aggregate-consumption Euler equations imply intertemporal elasticities of substitution near zero [Hall (1988), Campbell and Mankiw (1989)]. Unfortunately, the assumptions required for valid aggregation are extremely restrictive [Deaton (1992)], and there is some evidence that aggregation

leads to quantitatively significant biases in practice [Attanasio and Weber (1993)]. Most household-level studies imply that intertemporal elasticities of substitution are significantly greater than zero (Leontief preferences) and less than unity (Cobb–Douglas preferences), but estimates vary considerably within this range [see, e.g., Shapiro (1984), Zeldes (1989), Runkle (1991), Lawrance (1991), Dynan (1993), Attanasio and Weber (1993, 1995) and Attanasio and Browning (1995)]. Though the use of household panel data avoids the aggregation problems mentioned above, it necessitates other compromises. Panels are typically short, and data are often available only for isolated components of consumption (e.g. food)[35]. Microeconomic studies also frequently rely on variation in after-tax rates of return arising from cross-sectional differences in marginal tax rates, even though this variation is plausibly related to other pertinent household characteristics (e.g., factors explaining differences in wealth and income).

In the current context, it is also important to emphasize that one cannot infer the interest elasticity of saving directly from estimates of an Euler equation. Though Equation (26) provides information on the *shape* of the consumption profile, it does not tie down the *level* of consumption. This statement requires some clarification. In the simplest life-cycle models, the present discounted values of consumption and lifetime resources must be identical; consequently, one *can* infer the level of consumption, and thereby deduce the interest elasticity of saving, from the shape of the consumption profile. However, when one adds uncertainty, the intertemporal budget constraint becomes considerably more complex, and when one adds bequests (either intentional or accidental), the present discounted value of consumption need not equal the present discounted value of lifetime resources. In such models, the level of consumption is not recoverable from the shape of the consumption profile, and depends instead on a broader range of factors. Even fixing the parameters of the Euler equation, the implied interest elasticity of saving is sensitive to assumptions concerning bequest motives, the variability of income and expenses, risk aversion, and the prevalence of liquidity constraints (see Section 2)[36]. Similar statements hold for the welfare effects of alternative tax policies.

4. Evidence on responses to tax-deferred savings accounts

The existing literature on tax-deferred savings accounts focuses primarily (though not exclusively) on Individual Retirement Accounts (IRAs) and 401(k)s in the United

[35] Attanasio and Weber (1995) address these problems by constructing a longer, synthetic panel using the Consumer Expenditure Surveys, which contain more comprehensive consumption data.

[36] Many of these parameters are difficult to estimate. For example, though the constant term in the Euler equation depends on the pure rate of time preference (ρ), it also depends on risk aversion and the variability of consumption (through μ). One cannot separately identify ρ and μ without further assumptions.

States. A large branch of this literature attempts to measure a direct effect: all else equal, how much less would contributors have saved had these programs not existed? This question is the focus of the current section[37]. Tax-deferred accounts may also affect saving indirectly, for example by displacing other types of pensions. I consider the available evidence on some indirect effects in Section 5.

4.1. Individual Retirement Accounts

The US government first permitted individuals without pensions to open Individual Retirement Accounts (IRAs) in 1974. These accounts featured tax-deductible contributions up to a fixed limit, tax-free accumulation, taxation of principal and interest on withdrawal, and penalties for early withdrawal. Congress extended eligibility to all workers in 1981, and raised annual contribution limits to $2000 for an individual worker, or $2250 for a married couple with one earner. The Tax Reform Act of 1986 restricted eligibility for *deductible* contributions, based on adjusted gross income (AGI). Deductibility was phased out for AGI between $40 000 and $50 000 for joint filers, and between $25 000 and $35 000 for single filers. Individuals with higher levels of AGI remained eligible to make non-deductible contributions up to the same annual limits, and continued to benefit from tax-free accumulation. Beginning in January, 1998, taxpayers could also make contributions to Roth IRAs, which feature non-deductible contributions up to the same fixed limit, tax-free accumulation, tax-free withdrawal of contributions and earnings, and penalty-free early withdrawal of contributions.

Prior to the Tax Reform Act of 1986, IRAs had become quite popular. Annual contributions grew from roughly $5 billion in 1981 to roughly $38 billion in 1986, representing approximately 20 percent of personal saving. Contributions plummeted after 1986, falling to less than $10 billion in 1990. While it is indisputable that the flow of saving through IRAs was substantial, there is considerable controversy concerning to the extent to which this flow represented new saving. The existing evidence on the efficacy of IRAs falls into five general categories.

4.1.1. Direct survey evidence

One approach to measuring the effect of IRAs on saving is simply to ask people how they funded their contributions. In one such survey [Johnson (1985)], about half of respondents said that they would have saved the money anyway, about 10 percent said that they would have spent all of it, and about 40 percent said that they would have saved some and spent some. Johnson concludes that, on average, individuals reduced consumption by roughly 32 cents to fund each dollar of IRA saving.

[37] Other useful surveys include Hubbard and Skinner (1996), Poterba, Venti and Wise (1996a,b), Engen, Gale and Scholz (1996a,b) and Bernheim (1997c).

Evidence of this type suffers from a variety of problems. The relevant survey question asks individuals to imagine what they would have done in a counterfactual and purely hypothetical situation. Respondents may not think very hard about the hypothetical. If they think about it, they may assess the costs and benefits of various decisions differently than they would have in practice. They may accurately report what their *intention* would have been in the hypothetical situation, but actions and intentions do not always coincide. They may also misrepresent their probable intentions in the hypothetical situation, particularly if they believe that some answer is more "virtuous", or if they think that the interviewer is looking for a particular response.

4.1.2. Evidence on the frequency of limit contributions

Historically, roughly seventy percent of all IRA contributors save at exactly the contribution limit [Burman, Cordes and Ozanne (1990), Gravelle (1991b)]. Some analysts contend that the IRA program does not encourage thrift among these limit contributors because it fails to reduce their implicit price of future consumption, relative to current consumption, on the margin (the substitution effect). If this premise is valid, then IRAs may actually reduce saving through an inframarginal income effect (see Section 2.1.1).

As a matter of theory, there is no compelling reason to accept the premise mentioned in the previous paragraph. In the most basic life cycle model, individuals always wish to surpass contribution limits even if this requires them to borrow or shift assets (see Section 2.1.1). Consequently, binding contribution limits do indeed reflect the absence of a substitution effect, and the impact of IRAs is dominated by the inframarginal income effect. However, this basic model also has the counterfactual implication that *all* individuals should make the maximum allowable contribution. In fact, many households do not contribute at all, 30 percent of contributors do not reach the limit, and 70 percent of contributors fall short of the limit at least once over a three-year period [Hubbard and Skinner (1996)].

To account for non-limit contributors in the context of the life-cycle model, one must assume that individuals face borrowing restrictions and value liquidity (see sections 2.2.2 and 2.2.3). In that case, the substitution effect is certainly present for non-limit contributors. More to the point, it may also be present for limit contributors[38]. IRAs may encourage some individuals to increase their long-term, illiquid saving until they reach the contribution limit, and this increase may come at the expense of consumption, rather than liquid saving.

[38] Since the existence of the contribution limit induces a kink in the individual's budget constraint, it is simply incorrect to argue that a limit contributor's marginal rate of return is the same as in the absence of IRAs. Rather, the marginal rate of return on tax-deferred investments is not well-defined at the kink. The marginal rate of return on long-term, illiquid investments is well-defined (and unaffected by the existence of the IRA) only if the limit contributor has additional investments of this kind.

If one credits behavioral alternatives to the life-cycle hypothesis (Section 2.3), then evidence on the frequency of limit contributors is even less pertinent. Contribution limits may encourage saving by validating specific saving targets. IRAs may increase awareness of the need for retirement saving, or enhance efforts to impose self-discipline. Even if IRAs do not stimulate *inflows* into households' long-term savings, they may deter *outflows* [Thaler (1994)].

4.1.3. Correlations between IRA and non-IRA saving

A number of authors have attempted to measure the effects of IRAs on saving through more rigorous econometric analysis. Most of these studies have, with varying degrees of sophistication, examined the underlying correlations between IRA and non-IRA saving activity.

Before describing these studies, it is useful to begin by describing an ideal experiment for assessing the effects of IRAs. The contrast between the ideal data and the available data explains why the measurement of IRA effects has proven so difficult. Imagine that we are given some large sample of individuals, and that we randomly partition this sample into two subsamples. We treat the individuals in these subsamples exactly the same in all respects (identical initial assets, wages, fringe benefits, working conditions, and so forth), but we permit the individuals in one subsample to contribute to IRAs (the "experimental" group), while withholding this opportunity from the other subsample (the "control" group). In this way, we create exogenous variation in IRA eligibility. We then compare the total saving of individuals in the two subsamples to determine the effects of IRAs.

Unfortunately, between 1982 and 1986, there is no exogenous variation in IRA eligibility. Instead, we observe variations in participation. One could imagine attempting to mimic the ideal experiment by using this variation to identify new "experimental" and "control" groups, in effect asking whether the saving or assets of IRA contributors are higher than, lower than, or the same as the saving or assets of non-contributors. Evidence based on this approach reveals that IRA contributors do not save less in other forms than non-contributors; in fact, they save a good deal more [see e.g. Hubbard (1984)[39]]. Unfortunately, this finding tells us very little about the extent to which IRAs displace other saving. Some households save a lot, while some save little. This is presumably attributable to differences in preferences. Since the decision to contribute is endogenous, contributors probably consist of households with stronger preferences for saving. Therefore, one should not be surprised to discover that those who contribute to IRA accounts also save more in other forms than those who choose not to contribute.

[39] Hubbard's (1984) data are drawn from the 1979 President's Commission on Pension Policy, and therefore include some non-contributors who were ineligible for IRAs. Thus, the sample-selection problem discussed in the text is perhaps less pronounced than for estimates based on data collected between 1982 and 1986.

In principle, one solution to this problem would be to identify some exogenous variation in IRA contributions that is unrelated to preferences towards saving. One could then use instrumental variables to estimate a specification explaining non-IRA saving or total saving as a function of IRA saving. Since eligibility was universal from 1982 to 1986, a potential source for this variation is difficult to imagine, let alone measure.

Rather than attempt to identify an instrumental variable, the literature has proceeded by re-examining the relation between IRA saving and non-IRA saving, controlling for initial wealth. This procedure is based on the assumption that two individuals with the same initial wealth must have the same underlying preferences towards saving; thus, the source of the spurious upward bias between IRA saving and total saving is supposedly removed. This approach has been followed in a study by Feenberg and Skinner (1989) and a series of studies by Venti and Wise (1986–1988, 1990–1992). Analysis of a variety of data sources (including the Michigan Tax Panel, the Survey of Consumer Finances, the Consumer Expenditure Surveys, and the Survey of Income and Program Participation) uniformly demonstrate that total saving is positively correlated with IRA saving, even when one controls for initial wealth. The conditional correlation between IRA saving and non-IRA saving is typically non-negative. Some studies have interpreted these findings as indicating that IRA contributions are new wealth.

The central problem with this strategy is that initial wealth may be a relatively poor control for an individual's current underlying disposition toward saving. One problem is that wealth varies for reasons unrelated to tastes for saving (such as the receipt of unexpected inheritances). Another difficulty is that an individual's disposition to save may change through time due to fluctuations in income, household composition, perceived needs, or other factors; thus, the individual's disposition to save during any time period may differ from the dispositions that led to the accumulation of initial wealth at the start of the period. Even if wealth were perfectly correlated with the relevant aspects of tastes, it is well known that asset values are measured with a great deal of error. Any residual unobserved variation in the current inclination to save that is left after controlling for initial wealth will continue to bias the correlation between IRA saving and non-IRA saving upward: those who, for unobserved reasons, are inclined to save more overall will probably save more in both forms.

The underlying econometric justification for this procedure is also suspect. Even if it were possible to control perfectly for all aspects of tastes that determine non-IRA saving, this would not allow one to calculate the extent to which IRA contributions displace other saving, unless one could identify some significant exogenous variation in IRA contributions independent of tastes for saving. But with universal eligibility, it is hard to imagine any significant factor that would have affected IRA saving without also directly affecting non-IRA saving. If there is no source of exogenous variation in contributions, the relation of interest is presumably not identified.

In some of their work, Venti and Wise also place additional structure on the data. Specifically, they estimate the parameters of a model in which an individual maximizes a utility function defined over consumption, IRA saving, and non-IRA

saving. The specification allows for a range of elasticities of substitution between the two forms of saving. Based on estimates of this model, Venti and Wise conclude that IRA contributions represented new saving, in the sense that they were funded almost entirely by reductions in consumption and income taxes.

The low estimates of the substitution parameter that emerge from estimation of the Venti–Wise model appear to be driven by two considerations. The first consideration is the non-negative correlation (noted above) between IRA saving and non-IRA saving, conditional on initial wealth (which appears in the Venti–Wise model through the budget constraint). For reasons that I have already discussed, this correlation is probably a poor barometer for the true degree of substitutability.

The second consideration has to do with a technical feature of the model. As formulated, the model implies that, if IRA saving and non-IRA saving are perfect substitutes, then no individual would be willing to engage in non-IRA saving until reaching the IRA contribution limit. Since this prediction is manifestly false (many individuals who saved something did not contribute to IRAs), Venti and Wise's estimation strategy automatically guarantees the result that the two forms of saving are imperfect substitutes. This inference is unwarranted. Although it is evident that IRA saving and non-IRA saving must not be perfect substitutes for savers who do not contribute to IRAs (perhaps due to differences in liquidity), it does not follow that these two forms of saving are poor substitutes for individuals who do contribute to IRAs. On the contrary, one could easily imagine that, among IRA contributors, IRAs are quite good substitutes for other saving. This could occur if, for example, IRA contributors tend to save a lot in all forms, and are therefore relatively unconcerned (on the margin) about liquidity.

Gale and Scholz (1994b) estimate an alternative econometric model, in which they permit the parameters of the saving relation to vary according to whether or not an individual is an IRA contributor. This is intended to capture the possibility that those who do not contribute to IRAs may have different attitudes towards IRA and non-IRA saving than those who do contribute. In this way, Gale and Scholz avoid the automatic bias towards low substitution that is present in the analysis of Venti and Wise.

Intuitively, Gale and Scholz identify the degree of substitution between IRA and non-IRA saving as follows. Suppose we measure the marginal propensity to save (out of income) in IRAs ($\text{MPS}_{I,N}$), and the marginal propensity to save in other forms ($\text{MPS}_{O,N}$) for non-limit contributors, as well as the marginal propensity to save in other forms ($\text{MPS}_{O,L}$) for limit contributors. If all IRA saving is new saving, then we should find $\text{MPS}_{O,L} = \text{MPS}_{O,N}$. On the other hand, if IRA saving simply displaces other saving dollar-for-dollar, we would expect to find $\text{MPS}_{O,L} = \text{MPS}_{O,N} + \text{MPS}_{I,N}$. On the basis of this kind of comparison, Gale and Scholz conclude that a negligible fraction of IRA contributions represent new saving[40].

[40] According to Poterba, Venti and Wise (1996a), this central finding of Gale and Scholz is sensitive to changes in the specification and in the criteria used for selecting the sample.

The analysis of Gale and Scholz suggests that the conclusions of Venti and Wise are sensitive to assumptions about the nature and distribution of unobserved preferences. This does not imply, however, that their particular procedure generates reliable estimates of the extent to which IRAs substitute for other forms of saving. Identification of the Gale–Scholz model depends on the assumption that all IRA contributors have the same preferences towards saving, conditional on a list of covariates, regardless of whether they are limit contributors. There is an obvious tension between this assumption and the motivating premise of their analysis, which holds that attitudes towards saving differ according to IRA participation status even when conditioned on the same list of covariates.

To understand the potential bias resulting from the Gale–Scholz homogeneity assumption, consider the following illustrative example. Suppose that there are three types of savers, with (respectively) low, medium, and high inclinations to save. Those with greater inclinations to save are assumed to have larger average and marginal saving propensities. Low savers never contribute to IRAs, and are therefore of no further interest to us. As long as moderate savers are not constrained by the IRA contribution limit, they save 5 cents out of each dollar in IRAs, and 5 cents in other forms. If they are constrained by the contribution limit, they still save 5 cents out of each dollar in other forms. As long as high savers are not constrained by the IRA contribution limit, they save 10 cents out of each dollar in IRAs, and 10 cents in other forms. If they are constrained by the contribution limit, they still save 10 cents out of each dollar in other forms. Our final assumption is that all moderate savers end up contributing less than the contribution limit, while all high savers turn out to be limit contributors.

Note that, in this example, all IRA contributions represent new saving. However, applying the Gale–Scholz procedure, one would calculate that $\text{MPS}_{O,L} = 0.10 = 0.05 + 0.05 = \text{MPS}_{O,N} + \text{MPS}_{I,N}$, and infer incorrectly than IRA saving completely displaces other forms of saving. I have constructed this particular example to demonstrate that the bias could be quite large. Obviously, hypothetical examples cannot establish the magnitude of the actual bias. However, the principle (and therefore the direction of the bias) generalizes: heterogeneity among those who contribute to IRAs typically implies that those who contribute more (and who therefore have higher average propensities to save) will also tend to have higher marginal propensities to save. As a result, the data will appear to show that the marginal propensity to save in forms other than IRAs rises as contributions pass the allowable limit. But this is precisely the pattern that Gale and Scholz would interpret as evidence of displacement.

Some authors argue that correlations between IRA saving and non-IRA saving are particularly informative for new contributors. Using 1984 and 1985 data from the Survey of Income and Program Participation (SIPP), Venti and Wise (1995a) demonstrate that the inception of IRA contributions for a household does not coincide with a significant decline in other financial assets. They interpret this to mean that even new contributors engage in very little asset shifting to fund contributions, and that these contributions must therefore represent new saving. Yet the observed patterns

do not rule out the possibility that many new contributors were simply people with positive current shocks to saving, in which case these individual would have increased non-IRA savings in the absence of IRAs. Consequently, the evidence is consistent with significant asset shifting.

Attanasio and De Leire (1994) undertake a similar exercise, but suggest that it is appropriate to evaluate the behavior of new contributors treating old contributors as a control group. If new contributions come from consumption and if new and old contributors have similar preferences, then (it is argued) new contributors should exhibit slower consumption growth, and essentially the same growth in non-IRA assets, as old contributors. In contrast, if new contributions come from saving, then new contributors should exhibit the same growth in consumption, but slower growth in non-IRA assets than old contributors. The authors implement this test using the Consumer Expenditure Surveys, and find the second of these patterns. They conclude that IRA contributions primarily reflect asset reshuffling, rather than new saving.

Unfortunately, there does not appear to be any compelling justification for using old IRA contributors as a control group. It is natural to conjecture that old contributors opened IRA accounts earlier than new contributors because they have stronger innate predispositions to save. Obviously, this would account for their higher rates of saving. In principle, Attanasio and De Leire rule this possibility out by showing that consumption does not grow more rapidly for old contributors than for new contributors[41]. However, as a practical matter, consumption growth rates appear to be poor indicators of intrinsic thrift [see Bernheim, Skinner and Weinberg (2001)].

Even if old contributors were a valid control group, Attanasio and De Leire's inference would not follow. If new and old contributors have similar preferences and if IRA contributions reflect asset shifting for both groups, then one should not observe any systematic differences in either saving or consumption, contrary to the authors' findings. The observed pattern would instead suggest that contributions among new participants reflect asset shifting, while contributions among old participants represent new saving.

4.1.4. Exogenous changes in eligibility

Another possible approach to mimicking the ideal experiment is to exploit the exogenous variation in IRA eligibility that existed prior to 1982 and after 1986. For example, one could imagine estimating a regression explaining non-IRA saving as a function of IRA contributions using eligibility as an instrument, or directly as a function of eligibility. There are two problems with this suggestion; one is conceptual, the other practical. Conceptually, a problem arises because, in contrast

[41] If differences in saving result from differences in the pure rate of time preference and if the intertemporal elasticity of substitution is positive, then, under the LCH, those who prefer to save more would experience more rapid consumption growth.

to the ideal experiment, IRA eligibility was non-random. Eligibility was triggered by the absence of pension coverage prior to 1982, and by a combination of pension coverage and AGI after 1986. Since both pension coverage and income are potentially important determinants of household saving, concerns about correlations with underlying preferences are still present. The practical problem arises because, with certain data sources, eligibility is difficult to assess. Information on pension coverage is sometimes unavailable, incomplete or inaccurate, and one must extrapolate AGI from income.

The concern that IRA eligibility (prior to 1982 or after 1986) might have been correlated with preferences towards saving leads to a slightly more sophisticated suggestion. If the heterogeneity in preferences is captured by an individual-specific fixed effect, then it should be possible to eliminate this heterogeneity by differencing saving. One can then relate changes in saving to changes in eligibility, which differed across individuals both in 1982 and 1987. The impact of IRAs is then, in effect, inferred from differences in differences. For example, using panel data that crosses 1982, one attempts to determine whether those who became eligible for IRAs increased their saving by more than those who remained eligible.

This is the general approach taken in Joines and Manegold (1995) and Engen, Gale and Scholz (1994). Both of these studies make use of the IRS/University of Michigan Tax Panel. Unfortunately, this data set contains no information on pension coverage, and therefore provides no way to measure IRA eligibility prior to 1982. Of course, individuals who contributed to IRAs prior to 1982 must have been eligible. Joines and Manegold therefore propose using this as the control group. By defining the control group in this way, they tend to select individuals who have the highest predispositions to save among the eligible population. To counteract this selection effect, they use as their experimental group a sample of individuals who also contributed to IRAs (and therefore who also have high predispositions to save), but who began to contribute after 1982. While this experimental group includes some individuals who were eligible prior to 1982, it also includes many individuals who became eligible as of 1982. Therefore, on average, allowable contributions increased by a larger amount for members of the experimental group than for members of the control group. Both studies nevertheless demonstrate that there is relatively little difference between the changes in saving across 1982 for the experimental and control groups. Their preferred estimates suggest that IRAs had a moderate effect on saving (raising the contribution limit by one dollar raises saving by less than 30 cents).

One difficulty encountered by Joines–Manegold and Engen–Gale–Scholz is that the IRS/University of Michigan Tax Panel does not contain measures of either saving or wealth. The authors are compelled to impute wealth from dividend and interest income. They then difference estimated wealth to obtain a measure of saving. This variable is the focus of their differences-in-differences analysis. Thus, their key results are based on third differences (twice across time and once across subgroups) of an imputed variable. One must seriously question how much "news" is left over after these

operations. Not surprisingly, the authors obtain a wide range of estimates, and the key effects are generally estimated with large standard errors.

The selection criteria used to construct the control subgroup and the experimental subgroup are also potentially problematic. It is doubtful that these groups have comparable characteristics or similar dispositions to save. The differences-in-differences procedure is ostensibly designed to handle this problem, since it removes the fixed effect for each group. However, the validity of this solution depends critically on two assumptions: that tastes enter the saving equation additively, and that tastes do not affect the size of the response to a given change in the policy variable. In this context, these assumption are objectionable.

To further explore this point, suppose that the saving of group i at time t is given by the following equation:

$$s_{i,t} = \mu_i + \alpha_t + \eta_i M_{it}, \tag{27}$$

where μ_i and η_i are fixed group-specific coefficients, α_t is a time effect, and M_{it} is the IRA contribution limit applicable to this group. One would expect μ_i and η_i to be positively correlated, since higher savers are more likely to respond to an increase in the contribution limit. The differences-in-differences estimator is then

$$\Delta s_{e\,t} - \Delta s_{c\,t} = [\eta_e \Delta M_{e\,t} - \eta_c \Delta M_{c\,t}] \tag{28}$$

(where "e" indicates the experimental group, and "c" indexes the control group). Note that one will correctly estimate the effect of the policy change on the experimental group as long as $\eta_e = \eta_c$ (i.e. if there is no heterogeneity in the response to a given change in policy), or if $\Delta M_{c\,t} = 0$ (i.e. the control group does not experience a change in the policy variable)[42]. In this instance, neither condition applies: it is likely that heterogeneity in preferences towards saving (as reflected in η_i) remains, and contribution limits were raised for the control group (albeit to a lesser extent than for the experimental group, so that $\Delta M_{e\,t} > \Delta M_{c\,t} > 0$).

The resulting bias in the estimates depends on whether the control group is innately more inclined to save or less inclined to save than the experimental group. Suppose for the moment the control group consists of particularly high savers, so that $\eta_c > \eta_e$. Then the sign of the differences-in-differences estimator becomes ambiguous, even if an increase in the contribution limit actually stimulates saving for both groups. To take an example, if a \$2000 increase in the contribution limit induces a \$1000 increase

[42] Even if one of these conditions were satisfied, one would still obtain a biased estimate of η_e in practice. Recall that the experimental group is contaminated by the inclusion of households that were eligible to make IRA contributions prior to 1982, and that therefore properly belong in the control group. By ignoring this problem, these studies overstate the average change in the contribution limit for members of the experimental group ($\Delta M_{e\,t}$). If the true value of η_e is positive, this implies that the estimated value of η_e is biased downward [since it equals $(\Delta s_{e\,t} - \Delta s_{c\,t})/(\Delta M_{e\,t} - \Delta M_{c\,t})$].

in the average IRA saving of the control group and a $250 increase in the average saving of the experimental group (because the control group largely consists of more highly motivated savers), then a $500 increase in the contribution limit for the control group (e.g., from $1500 to $2000) and a $2000 change in the contribution limit for the experimental group (e.g., from $0 to $2000) will have the same total effect on saving ($250).

Unfortunately, with the available data, it is impossible to test whether the control group is more or less predisposed to undertake long-term saving than the experimental group. However, the following is suggestive. Prior to 1982, only a tiny fraction of those eligible for IRAs actually made contributions. While these individuals had one characteristic that might be indicative of a predisposition for low saving (no employer pension), they were nevertheless a very highly selected subset of this population. The fact that they both discovered and took advantage of a little-known IRA provision suggests that they may have been exceptionally motivated to save for retirement. In contrast, since a much larger segment of the population contributed to IRAs after 1982, and since IRAs were more widely publicized, the experimental group may be less highly selected. If this is the case, then the differences-in-differences estimator understates the true effect on saving of an increase in the IRA limit. Of course, if the opposite proposition is true, then the differences-in-differences estimator overstates the effect[43].

As is well known, the differences-in-differences estimator may go awry for other reasons as well. One obvious possibility is that other changes in the economic environment may have affected the two groups differently. Since the changes in IRA eligibility were accompanied by other significant tax changes, as well as a variety of important macroeconomic developments (including large changes in inflation and interest rates, as well as business-cycle effects), attributing the difference-in-difference of saving exclusively to relative changes in IRA eligibility is dicey.

Finally, it is important to realize that, under some of the behavioral alternatives to the LCH, the procedure used by Joines–Manegold and Engen–Gale–Scholz would be incapable of detecting certain kinds of links between IRAs and personal saving. Suppose, for example, that the expansion of the IRA program stimulated saving by enhancing awareness of retirement issues, creating peer-group effects, and triggering aggressive promotion of investment vehicles (see the discussion of the evidence on psychological effects, immediately below). If these developments affected members of the control group and the experimental group equally, the differences-in-differences estimator would falsely indicate no increase in saving.

[43] Engen, Gale and Scholz (1994) replicate Joines and Manegold's procedure, but also estimate a fixed-effects model using the full sample, treating all non-contributors prior to 1982 as if they were ineligible. In effect, this enlarges the Joines–Manegold experimental group by adding households that were eligible prior to 1982, but that never contributed to an IRA. This increases the likelihood that members of the experimental group are, on average, less inclined to save than members of the control group, and is therefore more likely to build in a bias against the hypothesis that IRAs added to total saving.

4.1.5. Evidence of psychological effects

The theoretical arguments that lead one to doubt the efficacy of IRAs are largely predicated on the view that saving is a consequence of rational and deliberate life-cycle planning. It is therefore possible to shed light on the key issue by asking whether other aspects of individuals' responses to IRAs are consistent with the predictions of standard life-cycle theory. If they are not consistent, then one should be very cautious about drawing inferences concerning the efficacy of IRAs from anything but the most direct evidence.

The literature identifies a number of patterns in IRA contributions that appear to be anomalous from the perspective of the standard model. The following four are particularly provocative.

First, it is difficult to account for the explosion of IRAs after 1982 and the collapse of IRA contributions after 1986, unless one credits the role of visibility and promotion [Long (1990), Venti and Wise (1992)]. Recall that only 1 percent of taxpayers made contributions to IRAs prior to 1982, despite the fact that roughly half were eligible to contribute up to $1500. This figure rose to 15 percent by 1986. Recall also that many individuals remained eligible to make deductible IRA contributions after 1986 (those with sufficiently low AGIs, or without pension coverage); moreover, all other individuals could still make non-deductible contributions and benefit from tax-free accumulation. Yet the fraction of taxpayers contributing to IRAs dropped to 4 percent by 1990. IRA contributions may have followed promotional activity (which exploded after 1982 and contracted after 1986) much more closely than actual economic incentives[44].

Second, there has been a pronounced tendency for individuals to delay their IRA contributions until the end of a tax year [Summers (1986)]. This is puzzling because minimization of tax liabilities requires taxpayers to make these contributions as early as possible. To some extent, the tendency to delay contributions may result from the desire to maintain liquidity throughout the tax year [Engen, Gale and Scholz (1994)]. But, even allowing for the potential importance of liquidity, it is difficult to explain why more IRA contributors (particularly those with significant non-IRA assets) do not at least make a series of smaller contributions during the course of the tax year [Bernheim (1994b)].

Third, individuals are significantly more likely to make IRA contributions if they owe the IRS money at the end of the tax year. Feenberg and Skinner (1989) interpret

[44] Engen, Gale and Scholz (1994) argue that IRA contributions may have declined after 1986 because of reductions in marginal tax rates and limits on deductibility. But unless one believes that the interest elasticity of saving is enormous, this could not have accounted for the magnitude of the decline in contributions. They also attribute the decline in IRA saving to the increased availability of 401(k)s and/or the possible depletion of non-IRA financial assets. There is little evidence to support this claim, and it is doubtful that either phenomenon can account for the sharpness of the decline in IRA contributions.

this as an indication that, psychologically, individuals would rather write a check to an IRA account than write a somewhat smaller check to the IRS. It is conceivable that this result could reflect spurious correlations of both underwithholding and IRA contributions with third factors, such as income, tax filing status, or asset holdings [Gravelle (1991b)]. However, the pattern is apparent even when Feenberg and Skinner include plausible controls for these factors.

Fourth, there is evidence of "focal point" saving. Engen, Gale and Scholz (1994) find that, among those who could have contributed more than $2000 but who contributed less than the limit, 47 percent contributed exactly $2000. This finding invites the interpretation that the well-publicized, "officially endorsed" $2000 figure created a focal target for saving, and that the very existence of this target may have influenced the behavior of many less serious savers (such as those contributing less than the limit)[45].

4.2. 401(k)s

Congress originally authorized employers to establish 401(k) plans in 1978, but this option remained unpopular until after the Treasury issued clarifying regulations in 1981. In many ways, 401(k)s are similar to IRAs: contributions are tax-deductible up to specified limits, the returns to investments are accumulated tax free, and there are restrictions on early withdrawals. There are also a number of important differences. Contribution limits are significantly higher and bind much less frequently. Consequently, there is general agreement that 401(k)s increase the marginal after-tax rate of return for most eligible households. This effect is often reinforced through provisions whereby employers match employee contributions. From a behavioral perspective, higher contribution limits may provide authoritative validation for higher saving targets. Moreover, 401(k)s may be more conducive to the exercise of self-discipline because contributions occur through regular payroll deductions rather than through discretionary deposits. Finally, since 401(k)s are organized around the workplace, they may create positive spillovers between employees (e.g., through conversations among employees and other "peer-group" effects).

[45] One alternative explanation for this phenomenon concerns transactions costs. While single-earner married couples could contribute up to $2250 per year, contributions in excess of $2000 would have required them to open a second account. A contribution of $250 might seem insufficient to justify the effort. However, it is important to bear in mind that the one-time costs of opening the account must be weighed not against the benefits of a single $250 contribution, but rather against the benefits of a $250 contribution that recurs for many years. Moreover, even among those with a $4000 limit, 38 percent of those contributing less than the limit contributed exactly $2000. Others have argued that the focal-point saving phenomenon results from bargaining among spouses with conflicting objectives [Burman, Cordes and Ozanne (1990)]. Yet it is hard to see how this would emerge in a formal model of household bargaining, without the introduction of significant transactions costs.

From the perspective of econometric modeling, one of the most salient differences between IRAs and 401(k)s is that eligibility for 401(k)s is determined at the level of the employer. This has two implications. First, at all points in time there is substantial variation in 401(k) eligibility across households. Second, at least some of the variation in eligibility (and therefore in contributions) arises from sources that are plausibly exogenous to the individual. These considerations make it easier *in principle* to identify the effects of 401(k)s.

Studying 401(k)s *in practice* is made considerably more difficult by the relative scarcity of good data. For example, none of the available waves of the Survey of Consumer Finances contains a clean measure of 401(k) eligibility. Of the standard public-use data sources, only the SIPP contains good information on eligibility, participation, and asset balances for 401(k)s. Unfortunately, the SIPP does not provide longitudinal information that is useful for studying these plans. The literature has therefore treated the SIPP as a series of three cross-sections (1984, 1987, and 1991). An additional limitation of these data is that 401(k) plan balances are not available in 1984. Taken together, these limitations seriously handicap efforts to measure the behavioral effects of 401(k)s. Nevertheless, the literature has developed and explored several estimation strategies that attempt to finesse these limitations.

4.2.1. Exploiting exogenous variation in eligibility

Imagine for the moment that each firm's decision to offer a 401(k) is completely random. Then 401(k)s would provide the perfect experimental setting for studying the effects of saving incentives. Eligibility is certainly not random, since it is demonstrably correlated with a variety of individual characteristics (such as income). But as long as variation in 401(k) eligibility is orthogonal to the unobserved individual characteristics that determine saving, the experiment is still a clean one.

Poterba, Venti and Wise (1994, 1995) proceed from the assumption that 401(k) eligibility is exogenous to the process that determines saving. Using the 1987 and 1991 waves of the SIPP, they demonstrate that, controlling for other relevant factors, eligibility is significantly correlated with median financial wealth. Indeed, eligibility has very little effect on median non-401(k) financial wealth. They interpret this finding as an indication that virtually all 401(k) contributions represented new saving.

The central problem with this procedure is that 401(k) eligibility is probably not exogenous. On the contrary, there are several reasons to suspect that eligibility would be significantly correlated with the underlying predisposition to save [Bernheim (1994c), Engen, Gale and Scholz (1994)]. First, employees with tastes for saving probably tend to gravitate towards jobs that provide good pension coverage, including 401(k)s. Second, employers frequently install 401(k) plans as a direct response to expressions of employee interest [Buck Consultants (1989)].

Asset ownership patterns are consistent with the view that 401(k) eligibility is endogenous. Specifically, differences in median financial assets between eligibles and non-eligibles are often several times as large as 401(k) balances for eligibles

[Poterba, Venti and Wise (1994), Engen, Gale and Scholz (1994), Bernheim and Garrett (2002)][46]. Either 401(k)s crowd-in other forms of saving at the implausible rate of four or five to one, or eligibility is strongly correlated with the innate inclination to save.

As in the case of IRAs, one could attempt to control for differences in tastes by using initial wealth as a taste proxy in a model explaining flow saving [see Bernheim and Garrett (2002)]. Unfortunately, as discussed in Section 4.1.3, observed wealth varies for many reasons that are unrelated to underlying tastes for saving. Consequently, some correlation between 401(k) eligibility and unobserved tastes for saving may remain, even when one conditions on initial wealth. This continues to bias the coefficient of interest[47].

4.2.2. Exploiting transitional effects

A second approach to measuring the effects of 401(k)s does not require one to assume that eligibility is exogenous. Instead, this approach exploits the fact that the legislative authorization for 401(k)s was relatively recent. To understand this approach, first imagine two idealized worlds, one in which 401(k)s have always been available, and one in which 401(k)s have never been available. Suppose for simplicity that each economy has converged to a steady state with a stable cross-sectional age–wealth profile. This profile may well be higher for the world in which 401(k)s have always been available, but this does not necessarily indicate that 401(k)s stimulate saving, since there may be other differences (such as tastes) between the two worlds. Now imagine a world in which 401(k)s have never been available in the past (so that this economy has also converged to a steady-state cross-sectional age–wealth profile), but where they are established unexpectedly as of some point in time (without any change in tastes). At that point, each individual departs from his or her initial wealth trajectory, and begins to move along some new wealth trajectory. Eventually, the cross-sectional age–wealth profile will converge to a new steady state. But during the transition period,

[46] This may seem inconsistent with the earlier statement that eligibility bears little relation to median non-401(k) financial wealth. Both statements are nevertheless accurate. The apparent anomaly occurs because median financial assets do not equal the sum of median 401(k) assets and median non-401(k) financial assets.

[47] Once one conditions on initial wealth, the direction of the bias is no longer clear. This is because the partial correlation between 401(k) eligibility and tastes for saving may be either positive or negative. To understand how it could be negative, imagine two individuals who are the same in all observable respects (including initial wealth), except that one is eligible for a 401(k), while the other is not. Suppose for the moment that 401(k)s actually stimulate saving to some unknown extent. It is very likely (due to the presence of high serial correlation in eligibility) that the eligible individual was also eligible in past years. Thus, without eligibility, this individual's initial wealth would have been lower than that of the individual who is actually ineligible. Consequently, under identical conditions (including eligibility), the eligible individual would have accumulated less wealth than the ineligible individual. This suggests that the ineligible individual is more inclined to save (given the observation that initial wealth is the same). If so, then assuming that 401(k)s stimulate saving, the estimated coefficient of eligibility would be biased downward.

if 401(k)s stimulate saving, this profile should begin to shift upwards relative to the profile from any world in which eligibility is unchanged.

In the ideal implementation of this strategy, one would identify a large group of workers who became eligible for 401(k)s relatively soon after the enabling legislation (say before 1984) and who remained eligible in all subsequent years, as well as a large group of workers who never became eligible for 401(k)s. One would then follow these same individuals through time, estimating cross-sectional age–wealth profiles for each group in each year. The relative amplitudes of these profiles in any particular year would prove nothing, since eligibility may be related to preferences. However, as time passes, the number of years of accumulated eligibility for the first group increases. Therefore, the cross-sectional age–wealth profiles for the eligible group should shift upward relative to the profile of the ineligible group.

Unfortunately, as mentioned above, good panel data on 401(k)s are not available. Poterba, Venti and Wise (1995) therefore implement this strategy for a series of cross-sections (1984, 1987, 1991) obtained from the SIPP. In each year, they compare the accumulated financial assets of those who are eligible for 401(k)s and those who are not eligible. The data unmistakably show the predicted upward shift in relative financial assets held by those who are eligible for 401(k)s. Indeed, there is no noticeable decline in the relative level of non-401(k) financial assets held by this group. According to the authors, this finding supports the hypothesis that individuals funded 401(k) contributions through a combination of reduced taxes and spending, and not by diverting funds that they would have saved in any event.

Of course, Poterba, Venti and Wise depart from the ideal strategy by using an unrelated sequence of cross-sections. It is important to consider how this affects their results. If successive cross-sections of eligibles (and ineligibles) are simply random draws from the same population of eligibles (ineligibles), then there is no problem. A problem only arises if the average innate inclination to save among eligibles (or ineligibles) changes systematically through time.

Since new workers became eligible for 401(k)s over time, it is virtually certain that some compositional changes occurred between the successive surveys used by Poterba, Venti and Wise[48]. Moreover, these compositional changes are necessarily problematic[49]. Recall that this methodology is motivated by the observation that the average innate inclination to save differs between eligibles and ineligibles. But then

[48] As noted by Engen and Gale (1997), some eligible workers also became ineligible over time, and the effects of this migration were most likely opposite those noted in the text. However, the predominant flow during this period was into the eligible pool.

[49] One obvious implication is that, as one moves forward in time by, say, four years, the average length of exposure to 401(k)s within the eligible population increases by less than four years. One can imagine cases in which this could create problems. For example, if 401(k)s pass through a period of sufficiently rapid growth, the average length of eligibility among eligibles could actually decline. However, under more plausible assumptions, the effect would simply be to slow the observed *rate* at which the assets profile of the eligible population shifts relative to the profile of the ineligible population.

the movement of individuals from the ineligible population into the eligible population must, of necessity, change the average innate inclination to save among eligibles, ineligibles, or both.

The duration and magnitude of the resulting bias depends on the characteristics of newly eligible workers. These individuals are probably systematically different from those who have been eligible for longer periods. The most motivated "serious" savers probably sought out employers who offered 401(k)s, or encouraged their existing employers to provide 401(k)s, relatively soon after these plans became available. Less motivated, "occasional" savers were probably less likely to seek out or agitate for 401(k)s, and more likely to drift into these plans slowly through time. Thus, as time passes, the eligible population becomes increasingly skewed towards less motivated savers. Bernheim (1994b) refers to this as the "dilution" effect[50]. It is likely that the dilution effect became more severe after 1986, when more demanding non-discrimination requirements were established for private pensions. Since dilution creates a *downward* shift in the estimated cross-sectional age–wealth profile of eligible workers, it has the potential to partially offset, completely offset, or even reverse any upward shift due to the behavioral effect of 401(k)s.

Were this the only effect of dilution, the direction of the resulting bias would be clear. However, migration of workers from the ineligible population into the eligible population also changes the composition of the ineligibles. Though newly eligible workers are probably less serious savers on average than those who have been eligible for longer periods, they are probably more serious savers on average than those who remain ineligible. Thus, as time passes, the ineligible population may *also* become increasingly skewed towards less motivated savers. Since this leads to a downward shift in the estimated cross-sectional age–wealth profile of ineligibles, it has the potential to create the spurious appearance that the profile for eligibles has shifted upward relative to the profile for ineligibles.

[50] To determine whether dilution occurs in practice, one can examine changes through time in the relations between 401(k) eligibility and variables that provide stable proxies for underlying tastes. One plausible proxy for the predisposition to save is ownership of an IRA. It is doubtful that this taste proxy is stable for the period of universal IRA eligibility (prior to 1987), since dilution probably affected the set of IRA participants in the same way that it affected the set of 401(k) participants. However, dilution of the IRA population probably declined significantly once eligibility for IRAs was restricted since the frequency of new participation fell dramatically. It is therefore plausible that IRA ownership is a stable taste proxy for the 1987–1991 period. Notably, the fraction of individuals eligible for 401(k)s who owned IRAs declined significantly (relative to ineligibles) between 1987 and 1991. This is a good indication of the dilution effect. Engen and Gale (1997) note that 401(k) participation rates have risen over time, and they assert that this is evidence of reverse dilution. Given the overall increase in 401(k) eligibility, it is more likely that rising participation results from other factors, such as increased familiarity with 401(k)s or the intensification of employer efforts to encourage participation. Since there is also a certain amount of "stickiness" to 401(k) participation decisions, one would also naturally expect participation rates to ratchet upward over time even without any change in the eligible population.

The net effect of dilution is theoretically ambiguous. If newly eligible workers are typical of the eligible population, then there will be a spurious downward shift in the cross-sectional age–wealth profile of the ineligible population, and no spurious shift in the profile of the eligible population. In that case, the approach would overstate the effects of 401(k)s. If newly eligible workers are typical of the ineligible population, then there will be a spurious downward shift in the cross-sectional age–wealth profile of the eligible population, and no spurious shift in the profile of the ineligible population. In that case, the approach would understate the effects of 401(k)s.

Engen, Gale and Scholz (1994) use the same approach as Poterba, Venti and Wise, but restrict attention to selected subgroups of the eligible and ineligible populations. Specifically, they compare cross-sectional age–wealth profiles for 401(k) *contributors* to profiles for individuals with IRAs who are ineligible for 401(k)s. The purpose of this strategy is to homogenize the unobserved preferences of eligibles and ineligibles by focusing in each instance on "high savers". The authors find that the cross-sectional age–financial wealth profile of 401(k) contributors actually shifted downward between 1987 and 1991, whereas the profile for the selected ineligibles shifted upward. They interpret this as indicating that 401(k)s did not increase personal saving. It is important to realize, however, that this approach continues to suffer from the dilution problem because it does not eliminate unobserved variation in tastes for saving. By changing the selection criteria used to define the samples, the authors have probably altered the nature and extent of dilution for the eligible and ineligible groups. Bernheim (1994b) argues that these changes reverse the direction of the dilution effect for the ineligibles, and thereby create a bias against the finding that 401(k)s increase saving[51].

It is also important to emphasize that Poterba, Venti and Wise focus exclusively on financial assets. This is a potential limitation, since 401(k)s may displace other forms of wealth. To evaluate the importance of this limitation, Engen and Gale (1997) make similar calculations using a broader definition of wealth. Their results indicate that mortgages grew and home equity fell in successive cross-sections for the 401(k)-eligible population (both IRA participants and IRA non-participants), resulting in smaller overall wealth growth than for the control groups. They interpret this finding as an indication that 401(k) saving was offset almost completely by larger mortgages.

Bernheim (1997c) questions the plausibility of the Engen–Gale results by arguing that, if 401(k)s do displace other saving, they are more likely to reduce the accumulation of financial assets than to encourage greater borrowing against homes. Concerns about plausibility are compounded by problems with Engen and Gale's evidence. While the absolute decline in home equity was greater for the 401(k)-eligible population than for the ineligible population, the 401(k)-eligible group started out with more housing wealth; the percentage decline was essentially identical for the two

[51] The argument is that there may have been migration out of IRA accounts after eligibility was restricted in 1986, and that those retaining their IRA accounts were presumably the most serious savers. This would lead to a spurious upward shift in the estimated cross-sectional age–wealth profile for ineligibles.

groups. This suggests that the phenomenon may be attributed to unrelated third factors. Naturally, the Engen–Gale procedure continues to suffer from the problems associated with dilution[52]. In addition, the results for total wealth are extremely imprecise. Engen and Gale typically cannot rule out (at conventional levels of confidence) the possibility that 401(k)s contributed significantly to total wealth accumulation. This raises the possibility that their finding might not be robust. Using the same data, Poterba, Venti and Wise (1996b) conclude that the timing of changes in mortgage debt and net home equity over time is inconsistent with a causal relationship between 401(k) contributions and mortgage debt. These conflicting findings are not easily reconciled.

Engen and Gale also point out that the cross-sectional age–wealth profiles of 401(k)-eligible renters did not shift upward relative to those of ineligible renters between 1987 and 1991[53]. This suggests that Poterba, Venti and Wise's central result may not be robust when one focuses on the population segment for which non-financial wealth is relatively unimportant. While these findings are thought-provoking, their proper interpretation is unclear. Renters as a whole are a peculiar group in that they save practically nothing to begin with [US Congressional Budget Office (1993)]. Those who are eligible for 401(k)s do accumulate significant financial assets (though significantly less than comparable homeowners); however, the median net worth of renters who are not eligible for 401(k)s is near zero. These observations have two implications. First, the sample of eligible renters appears to be more highly selected than the sample of eligible homeowners. As a result, eligibility for 401(k)s may be more strongly related to underlying tastes among renters than among homeowners. Sample selection biases and the associated effects of dilution should therefore play out differently in the two samples. It would not be surprising if eligible renters, being more highly selected to begin with, were subject to greater dilution with the passage of time. Second, sample-selection issues aside, the absence of significant wealth among ineligible renters can potentially invalidate the methodology used to draw inferences about the effects of 401(k)s. If economic forces were tending to depress saving by renters during the relevant time period, then the absence of a downward shift in the age–wealth profile for eligible renters would indicate that 401(k)s stimulated saving by this group. In theory, the Engen–Gale procedure would detect this by noting a downward shift in the age–wealth profile for ineligible renters. However, in practice, liquidity constraints would have prevented this downward shift from occurring.

[52] Alert to this issue, they attempt to control for unobserved preferences by including a measure of IRA participation status. Poterba, Venti and Wise employ a similar approach in some of their work. Unfortunately, this does not solve the problem. The trend in the probability that the typical 401(k) worker owns an IRA is properly regarded as a symptom of dilution, rather than as the source of the underlying problem [Bernheim (1994b)]. It is highly unlikely that this single binary variable adequately controls for the full variation of preferences towards saving among eligibles and ineligibles.

[53] There is some evidence of an upward shift between 1984 and 1991, but Engen and Gale argue that the 1987–1991 comparison is more reliable due to data limitations affecting the 1984 survey.

4.2.3. *Exploiting variation in matching rates*

Employers frequently match employee contributions to 401(k) plans, and there are substantial differences across employers both in the matching rates and in the amounts matched. The economic rewards associated with 401(k) saving therefore vary considerably, even among eligible workers. In principle, one could attempt to assess the effects of economic incentives (including taxes) on saving by exploiting this variation.

To date, relatively few studies have attempted to relate 401(k)-plan provisions, such as employer matches, to the choices of employees. Moreover, the existing studies focus exclusively on 401(k) contributions. Even if 401(k) contributions respond strongly to employer matching provisions, it is conceivable that this could reflect asset shifting rather than new saving. Thus, a high elasticity of contributions with respect to the match rate would not necessarily establish that individuals save more when the returns to saving are more generous. If, however, contributions do not rise with the match rate, then it seems unlikely that total saving would respond to changes in the after-tax rate of return.

The evidence on the effect of 401(k) match rates is mixed. Using survey data gathered by the General Accounting Office, Poterba, Venti and Wise (1992) conclude that the existence of a match rate is correlated with higher participation, but that the level of the match has little effect. Papke, Petersen and Poterba (1993) survey a small sample of firms and corroborate this finding. Papke (1992) analyzes data drawn from IRS Form 5500 filings, and finds that the effect of higher match rates is positive at low match rates, but negative at high match rates. Her results are somewhat sensitive to the introduction of fixed effects. Andrews (1992) studies household-level data from the May 1988 Current Population Survey, and concludes that, while the existence of a match increases participation, there is actually a negative relation between the match rate and contributions. Kusko, Poterba and Wilcox (1998) analyze employee-level data for a single company, and find that contributions and participation are relatively insensitive to changes in the matching rate through time. Scott (1994) argues that most of the negative results on the effects of matching provisions are attributable to the use of *ex post* rather than *ex ante* match rates. Using the 1985–1989 Employee Benefit Surveys (for which *ex ante* match rates are available), he finds some evidence that the size of the match matters; however, even Scott's results indicate that most of the effect is attributable to the existence of the match, rather than to its magnitude.

The evidence on match rates is therefore somewhat puzzling. Within the context of the traditional life-cycle hypothesis, it is surprising (though conceivable) that employees would respond so differently to match rates of 0% and 5%, but behave almost identically with match rates of 5% and 100%. Naturally, these findings could be spurious if the existence of a match is positively correlated with the underlying preferences for saving among employees. This would occur if, for example, high-saving workers sort themselves into plans with matches, or demand that their employers provide matches. It is, however, hard to understand why the same considerations would not induce a correlation between contributions and the size of the match. There is also

some reason to believe that matching provisions are adopted as remedial measures to stimulate contributions in instances where employees are predisposed against saving [Bernheim and Garrett (2002)]. In that case, the available results would understate the impact of a match.

The evidence on matching provisions is potentially reconcilable with alternative behavioral hypotheses. The availability of a match may focus employee attention on the 401(k) plan, authoritatively validate the importance of long-term saving objectives, undermine the resistance of impatient selves (due to the immediacy of the match), and provide additional impetus for establishing a private rule. Conceivably, these effects could emerge discontinuously with the introduction of a match, irrespective of its size.

4.3. General evidence from the US experience

In Section 4.2.2, I discussed the manner in which transitional phenomena generated by the relative novelty of 401(k)s have been used to assess their effects. More generally, one could regard the 1980s as a grand experiment with several different types of tax-favored accounts, and ask whether these accounts had the effect of shifting up the age–wealth profiles of entire cohorts. To take an example, if tax incentives were effective, then the typical individual reaching age 65 in, say, 1991 should have had more wealth than the typical individual reaching retirement in, say, 1984 (due to differences in years of eligibility for tax-favored saving).

Venti and Wise (1993) examine this hypothesis. Their analysis, which primarily relies on the SIPP, documents a substantial upward shift in financial asset profiles[54]. More recent cohorts have greater wealth at the same ages as older cohorts, and the difference is roughly equal to accumulated balances in 401(k)s and IRAs. While these patterns are interesting, it is potentially misleading to ascribe all differences in saving between cohorts to tax incentives. The same pattern could emerge if, for example, younger cohorts are wealthier on a lifetime basis.

4.4. Evidence from countries other than the United States

Although the existing literature has focused primarily on IRAs and 401(k)s, tax-deferred and/or subsidized savings accounts are not unique to the United States. Other programs include Canadian registered retirement savings plans, or RRSPs [Burbidge and Davies (1994)], and registered home-ownership savings plans, or RHOSPs [Engelhardt (1996)], British tax-exempt special savings accounts, or TESSAs, and personal equity plans, or PEPs [Banks and Blundell (1994)], the German *Vermögensbildungsgesetz* and *Bausparkassen* incentive programs [Börsch-Supan (1994)], the Italian treatment of life-insurance policies [Jappelli and Pagano (1994)],

[54] Engen, Gale and Scholz (1996b) identify several problems with the underlying data, and argue that the upward shift may be overstated.

Japanese *Maruyu* accounts [Ito and Kitamura (1994)], and French individual-savings plans, or PEPs, building society savings accounts, or CELs, and building-society savings plans, or PELs [Fougère (1994)].

Unfortunately, there is relatively little evidence on the effectiveness of these policies. A few studies use techniques similar to those discussed in sections 4.1 and 4.2 to analyze some of these programs; Venti and Wise (1995b) and Milligan (1998) study RRSPs, while Engelhardt (1996) examines RHOSPs. Others have attempted to deduce the effects of saving incentives from cross-country comparisons.

Although the generosity of the incentives embodied in tax-favored savings accounts differs significantly across countries, one cannot reliably infer the saving effects of these programs from simple cross-country correlations or regressions. If, for example, the political process is more favorable to the adoption of saving incentives in countries where voters care more about saving, then rates of saving will tend to be correlated with saving incentives even if these incentives have no effect on behavior.

A somewhat more subtile approach to international comparisons exploits the fact that different countries implemented their tax incentives at different points in time. This allows one to examine whether the saving rates of different countries converged or diverged when incentives were introduced. In this spirit, Carroll and Summers (1987) compare historical rates of saving for Canada and the United States. They demonstrate that these rates diverged when Canada expanded its system of Registered Retirement Saving Plans (RRSPs) during the mid-1970s. While this pattern is interesting, an inference of causality requires a leap of faith, particularly since there are other possible explanations for the increase in Canadian saving during this period. Moreover, the adoption of tax incentives in the USA did not result in measurable convergence between the two countries. More recent studies cast doubt on the hypothesis that tax-incentive programs account for relative movements of saving rates in the USA and Canada [see Sabelhaus (1997) and Burbidge, Fretz and Veall (1998)].

5. Evidence on other links between taxation and saving

Even if the interest elasticity of saving is low and households do not alter their behavior very much as a direct consequence of targeted tax incentives for saving, it might still be possible to influence personal saving through tax policy. In Section 2.3.1, I mentioned that non-neutralities in the tax system may encourage various kinds of third-party activities that have the potential to affect the level of personal saving. Specifically, non-neutralities may encourage employers to adopt various kinds of pension plans or to substitute one kind of plan for another, and may influence the activities of employers in the context of these plans. The tax system may also create incentives for corporations to save, or for the vendors of tax-favored financial vehicles to market and otherwise promote their products. In this section, I briefly summarize the evidence on each of these possibilities.

5.1. The size and scope of the pension system

Since pensions provide a tax-favored mechanism for compensating employees, tax policy may have played an important role in stimulating the development of the pension system. To assess the ultimate impact on personal saving, one must answer two questions. First, to what extent is the size and scope of the pension system responsive to changes in tax rates? Second, to what degree does pension saving displace other forms of personal saving? I consider these questions in turn.

5.1.1. Incentives for pension saving

It is indisputable that there is a substantial tax incentive for pension formation. Ippolito (1986) estimates that the optimum exploitation of opportunities to defer compensation through pensions can reduce lifetime tax liabilities by 20 to 40 percent. However, this does not imply that the growth of the pension system is exclusively, or even primarily attributable to the tax system. Pensions may enhance the productivity of the work force in a variety of ways. They may bond the workforce against union activity, voluntary job turnover, or poor job performance[55]. Employers may use defined-benefit plans to induce a desired pattern of retirement[56]. Mandatory pensions may also provide an effective device for overcoming the problems with adverse selection that characterize the market for private annuities[57]. Thus, it is conceivable that an extensive private pension system would exist even in the absence of tax incentives.

A number of studies provide empirical evidence on the relative importance of tax and non-tax determinants of pension coverage[58]. The central methodological problem in this literature is to identify an appropriate source of variation in marginal tax rates from which one can reliably infer tax effects. Time-series variation is primarily associated with a handful of significant tax reforms, and it is difficult to separate tax effects from confounding events. Since pension coverage can affect marginal tax rates, cross-sectional variation is potentially endogenous. To treat this problem, one must identify valid instrumental variables that are related to cross-sectional differences in marginal tax rates, but unrelated to the process that determines pension coverage.

Reagan and Turner (1995) adopt this approach, relying chiefly on cross-sectional variation in state income tax rates to identify the tax effect [see also Gentry and Peress

[55] See e.g., Ippolito (1985, 1986), Parsons (1986, 1995), Williamson (1992) and Allen, Clark and McDermed (1993).

[56] See e.g., Burkhauser (1979, 1980), Lazear (1984), Fields and Mitchell (1984), Ippolito (1986), Lazear and Moore (1988), Kotlikoff and Wise (1989), Stock and Wise (1990) and Quinn, Burkhauser and Myers (1990).

[57] See Ippolito (1986). Kotlikoff and Spivak (1981) discuss the nature of market failure in private annuity markets.

[58] Pertinent references includes Ippolito (1986), Bloom and Freeman (1992), Reagan and Turner (1995), Kruse (1995), Allen and Clark (1987), Woodbury and Bettinger (1991), Woodbury and Huang (1993), Clark and McDermed (1990), Feldstein (1994), Gentry and Peress (1994) and Gustman and Steinmeier (1995).

(1994)]. Their results imply that a one-percentage-point increase in marginal tax rates leads to a 0.4-percentage-point increase in pension coverage rates. The validity of this estimate presupposes the exogeneity of the state-income-tax variables. Conceivably, variation in tax rates across states could be related to differences in average income (which could in turn be correlated with the household's permanent income), or with other factors such as occupation or industry. Reagan and Turner attempt to control for these factors when explaining pension coverage, but their measure of permanent income is based on limited information, and their controls for occupation and industry are coarse.

5.1.2. Do pensions crowd out other personal saving?

The extent to which pensions displace other forms of personal saving probably depends on the characteristics of the pension. For our purposes, it is important to distinguish between employer-controlled pensions that provide the employee with no choice concerning the level of participation, and participant-controlled plans (such as 401(k)s) that permit the employee to determine contributions. I have already discussed the existing evidence on the extent to which contributions to participant-controlled plans crowd out other personal saving (Section 4.2). In this section, I focus on employer-controlled plans.

The existing literature contains more than a dozen studies that attempt to measure the degree of substitutability between pensions and other saving. The usual approach is to estimate a cross-sectional relation between either saving or wealth and some measure of pension coverage. The two earliest studies on this topic [Cagan (1965) and Katona (1965)] conclude that pensions actually crowd *in* other forms of saving. Cagan rationalizes this finding by arguing that pensions induce workers to recognize the need for retirement planning; he suggests that individuals may intensify their efforts to provide adequately for retirement because a pension renders this objective more feasible. Several subsequent studies corroborate the Cagan–Katona finding [Schoeplein (1970), Green (1981), Venti and Wise (1993), Bernheim and Scholz (1993a)]. More commonly, investigators have found either no effect, or a small effect [Munnell (1974), Kotlikoff (1979), Blinder, Gordon and Wise (1980), King and Dicks-Mireaux (1982), Diamond and Hausman (1984), Hubbard (1986), Wolff (1988), Samwick (1995), Gustman and Steinmeier (1998)]. Only a few studies have found substantial rates of crowding out [Munnell (1976), Dicks-Mireaux and King (1984), Avery, Elliehausen and Gustafson (1986), Gale (1995)], and most of these provide ranges of estimates that include relatively small effects. There is also some evidence that the rate of displacement rises with education [Bernheim and Scholz (1993b), Gale (1995)].

While there are many methodological concerns that bear on the reliability (both absolutely and relatively) of these various studies, three issues stand out as particularly salient. The first concerns the possibility that pension coverage is correlated with underlying tastes for saving. In contrast to the literature on 401(k)s, no existing study

has come to grips with this issue. The direction of the resulting bias is ambiguous[59]. The second issue concerns the measurement of compensation. For the most part, the studies listed above control for income, rather than total compensation (which would include the accrual of pension wealth). If the creation of a pension typically entails a shift in the form of compensation rather than incremental compensation, then this practice does not yield the appropriate displacement rate. Bernheim and Scholz (1993a) and Gale (1995) propose different solutions to this problem, and obtain very different results. The final issue concerns the definition of wealth. Although one can point to a number of exceptions, there is some tendency (as in the 401(k) literature) to find higher rates of displacement when one uses a broader measure of wealth. The issues here are similar to those mentioned in Section 4.2.2.

While the extent of crowding out is therefore not a settled issue, one is hard pressed to find convincing support in any study for the hypothesis that the rate of displacement is dollar-for-dollar. Indeed, there appears to be a significant likelihood that the true offset is much smaller. The importance of this finding becomes obvious when one considers that, between 1980 and 1990, the real change in pension assets exceeded the real change in national wealth by a wide margin [Shoven (1991)]. Thus, the effect of tax incentives on saving through the stimulation (or retardation) of pensions may be substantial, even if the rate of displacement is relatively high. Using estimates from the available literature, Engen and Gale (1996b) calculate that, following the replacement of the current income tax with a consumption tax, the reduction in saving due to changes in pensions could substantially or completely offset any increase in non-pension saving.

5.2. Employer-controlled pensions vs. participant-controlled pensions

In evaluating the extent to which 401(k)s contribute to personal saving (Section 4.2), I have abstracted from the degree to which these plans substitute for other pensions. If the rate of substitution is low, then policies that stimulate 401(k)s will tend to increase saving if and only if 401(k) contributions are not fully offset by reductions in non-pension saving. In contrast, if the rate of substitution is high, then policies that stimulate 401(k)s may increase or decrease saving, depending upon whether 401(k) contributions displace non-pension saving at (respectively) a lower or higher rate than other kinds of pensions.

Much has been written about the magnitude and probable causes of the shift from defined-benefit to defined-contribution pension plans in general, and to 401(k)s in particular [see, e.g., Parsons (1995), or Papke, Petersen and Poterba (1993), for

[59] Highly motivated savers may self-select into jobs with pension plans. But it is also conceivable that the workers who are most inclined to save, and who have the least problems with self-discipline, sort themselves into jobs that are covered by pension plans with the greatest discretion, such as 401(k)s. Those who are interested in saving, but who have problems with self-discipline, may prefer traditional employer-controlled plans.

selective reviews of this literature]. The existence of this shift does not, however, establish that 401(k)s have substituted for more traditional plans, since aggregate trends could in principle be driven by changes in the composition and organization of economic activity.

Papke, Petersen and Poterba (1993) examine data on individual firms, and conclude that wholesale replacement of existing plans (particularly defined-benefit plans) occurs in a minority of cases. While informative, this evidence does not resolve the central issue, since 401(k)s may displace other pension plans even if they do not directly replace these plans. For example, firms that adopt 401(k)s as supplementary plans may be less inclined to increase, and more inclined to decrease, the generosity of other pension plans. The available evidence also indicates that changes in industrial composition and the structure of firms cannot fully account for the aggregate shift to defined-contribution plans [see Clark and McDermed (1990), Gustman and Steinmeier (1992) and Kruse (1995)]. Since the unexplained component of the aggregate shift is large, it is possible that 401(k)s have substituted for other pension plans to a significant degree.

5.3. Taxation and corporate saving

Taxation affects corporate saving through two channels. First, an increase in the corporate tax rate reduces after-tax earnings. Unless corporations adjust dividends or share repurchases, retained earnings must fall. Second, both personal and corporate taxes may affect payout policy. For example, when the dividend tax rate rises relative to the effective tax rate for capital gains, corporations may pay smaller dividends.

There is a substantial body of theoretical and empirical work examining the effects of taxation on corporate payout and retention decisions. A review of this literature is beyond the scope of the current chapter; the interested reader should consult Alan Auerbach's chapter (19) in this Handbook. In this section, I consider the following related question: is it possible to stimulate total private saving through policies that encourage greater corporate saving?

In principle, private saving may be unresponsive to policies that successfully motivate corporations to save more. The reason is that households own corporations. When a corporation decides to pay dividends instead of retaining earnings, sophisticated shareholders should understand that the corporation is saving less on their behalf, and each shareholder should increase personal saving by an offsetting amount to reestablish his or her optimal life-cycle allocation.

Greater corporate saving might add to private saving if shareholders were liquidity constrained. In practice, however, share ownership is concentrated among higher-income individuals who are likely to have ample liquidity. At a minimum, these individuals have the option to borrow against or to sell their securities. Alternatively, shareholders might be irrational or myopic. One version of this view holds that investors suffer from a "bird-in-the-hand" fallacy: they believe that capital gains are transitory, and that income is more secure once it is actually received. Another version

of this view emphasizes the role of mental accounting: since dividend checks are cash-in-hand, they may be more spendable than an equivalent capital gain. Ultimately, the degree of substitutability between corporate saving and personal saving is an empirical question.

Early econometric studies of this issue involved the estimation of aggregated reduced form consumption functions. According to Feldstein (1973), for the USA, the marginal propensity to consume out of retained earnings is roughly two-thirds as large as the marginal propensity to consume out of disposable income [Feldstein and Fane (1973), obtain similar results for the UK]. Feldstein concludes that changes in private saving imperfectly offset changes in corporate saving, at the rate of 67 cents on the dollar. There are, however, alternative interpretations of Feldstein's findings. If retained earnings and disposable income have different stochastic properties (e.g., if the shocks to disposable income are more permanent than the shocks to retained earnings), then their coefficients in a reduced-form consumption function will differ. However, this implies nothing about the effects of shifting a deterministic dollar (or, for that matter, an income stream with fixed stochastic properties) between dividends and retained earnings. Feldstein's reduced-form consumption function approach also suffers from a variety of standard problems, including the potential endogeneity and/or imperfect measurement of key variables.

Poterba (1987, 1991) improves upon Feldstein's regressions in several respects[60]. Most notably, he uses a variable measuring the tax burden on dividends relative to capital gains as an instrument to treat the endogeneity of retained earnings[61]. To some extent, this also addresses the problem of interpretation mentioned above, since it yields a direct estimate of the effect on consumption of dollars shifted between retentions and payouts. Poterba finds that consumption rises significantly in response to tax changes that disfavor corporate saving. Notably, most of this effect occurs in the form of durable consumption, which is arguably another form of saving.

Poterba also examines the response of consumption to involuntary realizations of capital gains resulting from cash takeover transactions. In the absence of myopia or irrationality, one would expect shareholders to reinvest all of these gains. Yet Poterba's aggregate reduced-form consumption function estimates imply that investors increase consumption by about 60 cents for each dollar realized in such transactions. Once again, this effect is particularly strong for durable goods. These results appear to be driven by a limited set of events: personal saving declined sharply during the 1980s while takeover activity exploded. Since there are many other explanations for the decline in saving, the correlation could be coincidental.

[60] In addition to instrumenting retained earnings, he makes some important adjustments to the underlying data, distinguishes between durable and non-durable consumption, and estimates specifications in both levels and differences.

[61] One can criticize this choice of an instrument on the grounds that both tax rates belong in the consumption function regression.

Auerbach and Hassett (1991) adopt a much different approach to this same set of issues: they estimate aggregate-consumption Euler equations, and investigate whether changes in consumption are related to predictable changes in different components of income. The advantage of this approach is that it removes the informational effects that accompany unexpected changes in income and contaminate estimates of the marginal propensity to consume. Disadvantages include the usual range of objections to aggregate consumption Euler equations (see Section 3.2). Like others, Auerbach and Hassett find that consumption is sensitive to predictable changes in labor income[62]. In contrast, predictable changes in dividends and other forms of capital income have no effect on consumption. This finding undermines several hypotheses under which consumption would be sensitive to the division of corporate earnings between retentions and payouts. For example, it is inconsistent with the view that shareholders are liquidity constrained or more likely to spend cash-in-hand. It does not, however, rule out the possibility that individuals irrationally capitalize otherwise equivalent income streams of dividends and retained earnings at different rates, since changes in consumption would then occur only in response to unexpected changes in payout policy.

5.4. Other activities undertaken by employers

Aside from encouraging employers to provide various kinds of pensions, tax policy may also induce employers to engage in other activities that have the potential to influence saving. In some instances, this effect is indirect: by stimulating pensions, tax policy may also encourage activities that are complementary to pensions. In other cases, subtle features of the tax code may directly affect the activity in question.

Employer-based investment and retirement education is an example of an activity that is complementary to the provision of a pension plan. Tax policies that stimulate pensions in general, and especially participant-controlled plans, may also stimulate complementary educational initiatives [see Bernheim and Garrett (2002), Bayer, Bernheim and Scholz (1996) and Employee Benefit Research Institute (1995)]. Subtle features of the tax code, such as non-discrimination requirements, may also encourage employer-based retirement education more directly [in addition to the preceding references, see Garrett (1995)]. Generally, the impact of education is not subsumed in estimates of the relation between pensions and saving, since most of the growth of these offerings post-dates the most commonly used sources of data on household financial behavior.

There are a number of reasons to expect that retirement education might have an important effect on household saving. Various studies document low levels of financial literacy among adult Americans. This phenomenon is accompanied by an apparently widespread failure to appreciate financial vulnerabilities [Bernheim (1995)]. Although

[62] This is sometimes interpreted as evidence of liquidity constraints for those receiving labor income.

there is little direct evidence on the impact of educational programs, some recent studies conclude that employer-based offerings significantly stimulate both voluntary pension contributions and total household saving [see Bernheim and Garrett (2002), and Bayer, Bernheim and Scholz (1996), Bernheim (1998), Clark and Schieber (1998)]. Since the availability of employer-based retirement education may be correlated with employees' preferences, these studies potentially suffer from the usual kinds of sample-selection problems. However, there is some evidence that employers adopt these programs as remedial measures when employees have low predispositions to save (as indicated, for example, by low participation and contribution rates prior to adoption). In that case, the available evidence would understate the effects of educational interventions.

5.5. Marketing and promotion of financial products

The expansion of IRA eligibility to all taxpayers in 1981 was accompanied by a great deal of media fanfare. Perhaps more importantly, the existence of these retirement-saving vehicles created profit opportunities for financial institutions. Although the IRA tax incentive was targeted at households, it generated considerable impetus for private firms to promote saving through a blend of education and marketing. Similar phenomena occur in the context of other tax-deferred savings instruments, such as long-term life-insurance policies and variable annuities.

It is natural to wonder whether these promotional activities affect personal saving. Unfortunately, there is virtually no direct evidence on this issue. There are, however, two particularly interesting anecdotes. One concerns the introduction and subsequent scaling-back of IRAs, which I have discussed in Section 4.1.5. The other concerns experience with saving promotion in Japan [Central Council for Savings Promotion (1981)]. After World War II the Japanese government launched a national campaign to promote saving. Promotional activities included the organization of monthly seminars that extolled the virtues of saving and provided workers with financial guidance, the sponsorship of children's banks, the appointment of private citizens as savings promotion leaders, and the extensive dissemination of literature. While the Japanese rate of saving rose precipitously over the relevant time period, other factors were also at work, including the existence of strong tax incentives for saving, as well as various aspects of post-War reconstruction. One can therefore only speculate about the extent to which the increase in saving was attributable to promotion.

6. Concluding comments

From the discussion in the preceding sections, it is readily apparent that questions concerning taxation and saving have stimulated an enormous amount of research since the publication of Sandmo's (1985) survey in the original *Handbook of Public Economics*. This research has led to significant theoretical advances in our

understanding of the positive and normative implications of taxing the returns to saving, and has produced important contributions to our empirical knowledge of household behavior. Still, the critical analysis contained in this chapter underscores the limitations and shortcomings of the extant literature.

As an economist, one cannot review the voluminous literature on taxation and saving without being somewhat humbled by the enormous difficulty of learning anything useful about even the most basic empirical questions. Having been handed two grand "experiments" with tax policy during the 1980s (IRAs and 401(k)s), it would seem that we ought to have learned more, and to have achieved greater consensus, than we have. In our defense, it can be said that we have done our best with the information at our disposal. As I have mentioned at various points in this chapter, it is often easy to identify the kinds of data that would have allowed us to answer the pressing policy questions with much greater confidence. Unfortunately, we have had to make do with data that is, at best, a caricature of the ideal.

During the next decade, there will undoubtably be new experiments, and new opportunities to learn something useful about taxation and saving. The introduction of Roth IRAs in January 1998 provides one such opportunity, and I would expect this to generate a flurry of research activity once pertinent data become available. However, the prospects for significant advances in empirical methodology will be severely limited unless researchers have access to higher-quality data. When one thinks of the budgetary costs of tax incentives, and of what is at stake in terms of economic growth and efficiency, it seems a shame that ongoing, comprehensive, microeconomic data collection has been such a low social priority.

References

Aaron, H., and A.H. Munnell (1992), "Reassessing the role for wealth transfer taxes", National Tax Journal 45:119–144.

Ainslie, G. (1975), "Specious reward: a behavioral theory of impulsiveness and impulse control", Psychological Bulletin 82(4):463–496.

Ainslie, G. (1982), "A behavioral economic approach to the defense mechanisms: Freud's energy theory revisited", Social Science Information 21:735–779.

Ainslie, G. (1984), "Behavioral economics II: motivated, involuntary behavior", Social Science Information 23:47–78.

Ainslie, G. (1992), Picoeconomics (Cambridge University Press, Cambridge).

Aiyagari, S.R. (1995), "Optimal capital income taxation with incomplete markets, borrowing constraints, and constant discounting", Journal of Political Economy 103(6):1158–1175.

Allen, S.G., and R.L. Clark (1987), "Pensions and firm performance", in: Morris M. Kleiner et al., eds., Human Resources and the Performance of the Firm, Industrial Relations Research Association Series (Industrial Relations Research Association, Madison, WI) pp. 195–242.

Allen, S.G., R.L. Clark and A. McDermed (1993), "Pension bonding and lifetime jobs", Journal of Human Resources 28:463–481.

Altonji, J.G., F. Hayashi and L.J. Kotlikoff (1992), "Is the extended family altruistically linked? Direct tests using micro data", American Economic Review 82:1177–1198.

Andreoni, J. (1989), "Giving with impure altruism: applications to charity and Ricardian equivalence", Journal of Political Economy 97:1447–1458.

Andrews, E.S. (1992), "The growth and distribution of 401(k) plans", in: J. Turner and D. Beller, eds., Trends in Pensions 1992 (U.S. Department of Labor, Washington, D.C.) pp. 149–176.

Atkinson, A.B., and A. Sandmo (1980), "Welfare implications of the taxation of savings", The Economic Journal 90:529–549.

Attanasio, O.P. (1995), "The intertemporal allocation of consumption: theory and evidence", Carnegie-Rochester Conference Series on Public Policy 42:39–89.

Attanasio, O.P., and M. Browning (1995), "Consumption over the life cycle and over the business cycle", American Economic Review 85(5):1118–1137.

Attanasio, O.P., and T. De Leire (1994), "IRAs and household saving revisited: some new evidence", Working Paper No. 4900 (National Bureau of Economic Research).

Attanasio, O.P., and G. Weber (1993), "Consumption growth, the interest rate and aggregation", Review of Economic Studies 60(3):631–649.

Attanasio, O.P., and G. Weber (1995), "Is consumption growth consistent with intertemporal optimization? Evidence from the consumer expenditure survey", Journal of Political Economy 103(6):1121–1157.

Auerbach, A.J. (1979), "The optimal taxation of heterogeneous capital", Quarterly Journal of Economics 93:589–612.

Auerbach, A.J. (1996), "Tax reform, capital allocation, efficiency, and growth", in: H.J. Aaron and W.G. Gale, eds., Economic Effects of Fundamental Tax Reform (Brookings Institution, Washington, D.C.) pp. 29–73.

Auerbach, A.J. (2002), "Taxation and corporate financial policy", in: A.J. Auerbach and M. Feldstein, eds., Handbook of Public Economics, Vol. 3 (Elsevier, Amsterdam) ch. 19, this volume.

Auerbach, A.J., and K.A. Hassett (1991), "Corporate saving and shareholder consumption", in: B. Douglas Bernheim and John B. Shoven, eds., National Saving and Economic Performance (University of Chicago Press, Chicago) pp. 75–102.

Auerbach, A.J., and J.R. Hines Jr (2002), "Excess burden and optimal taxation", in: A.J. Auerbach and M. Feldstein, eds., Handbook of Public Economics, Vol. 3 (Elsevier, Amsterdam) ch. 21, this volume.

Auerbach, A.J., and L.J. Kotlikoff (1987), Dynamic Fiscal Policy (Cambridge University Press, New York).

Auerbach, A.J., L.J. Kotlikoff and J. Skinner (1983), "The efficiency gains from dynamic tax reform", International Economic Review 24:81–100.

Avery, R.B., G.E. Elliehausen and T.A. Gustafson (1986), "Pensions and social security in household portfolios: evidence from the 1983 Survey of Consumer Finances", in: F. Gerard Adams and Susan M. Wachter, eds., Savings and Capital Formation (Lexington Books, Lexington, MA).

Balcer, Y., and K. Judd (1987), "Effects of capital gains taxation on life-cycle investment and portfolio management", Journal of Finance 42(3):743–761.

Banks, J.S., and R. Blundell (1994), "Taxation and personal saving incentives in the United Kingdom", in: James M. Poterba, ed., Public Policies and Household Saving (University of Chicago Press, Chicago) pp. 57–80.

Barro, R. (1974), "Are government bonds net wealth?" Journal of Political Economy 82:1095–1117.

Barthold, T.A., and T. Ito (1992), "Bequest taxes and accumulation of household wealth: U.S.–Japan comparison", in: Takatoshi Ito and Anne O. Krueger, eds., The Political Economy of Tax Reform (University of Chicago Press, Chicago) pp. 235–290.

Bayer, P.J., B.D. Bernheim and J.K. Scholz (1996), "The effects of financial education in the workplace: evidence from a survey of employers", Working Paper 5655 (National Bureau of Economic Research).

Becker, G.S. (1974), "A theory of social interactions", Journal of Political Economy 82:1063–1093.

Benhabib, J., and A. Rustichini (1997), "Optimal taxes without commitment", Journal of Economic Theory 77(2):231–259.

Bernheim, B.D. (1987), "Ricardian equivalence: an evaluation of theory and evidence", Working Paper 2330 (National Bureau of Economic Research).

Bernheim, B.D. (1991), "How strong are bequest motives? Evidence based on estimates of the demand for life insurance and annuities", Journal of Political Economy 99:899–927.
Bernheim, B.D. (1994a), "Personal saving, information, and economic literacy: new directions for public policy", in: C.E. Walker, M. Bloomfield and M. Thorning, eds., Tax Policy for Economic Growth in the 1990s (American Council for Capital Formation, Washington, D.C.) pp. 53–78.
Bernheim, B.D. (1994b), "Comments and discussion", Brookings Papers on Economic Activity 1: 152–166.
Bernheim, B.D. (1994c), "Comment on chapters 4 and 5", in: David A. Wise, ed., Studies in the Economics of Aging (University of Chicago Press, Chicago) pp. 171–179.
Bernheim, B.D. (1995), "Do households appreciate their financial vulnerabilities? An analysis of actions, perceptions, and public policy", in: Tax Policy and Economic Growth (American Council for Capital Formation, Washington, D.C.) pp. 1–30.
Bernheim, B.D. (1997a), "Taxation and saving: a behavioral perspective", in: Proceedings of the Eighty-Ninth Annual Conference on Taxation, 1996 (National Tax Association, Washington, D.C.) pp. 28–36.
Bernheim, B.D. (1997b), "The Merrill Lynch baby boom retirement index: update '97", Mimeo (Merrill Lynch, Pierce, Fenner and Smith, Inc., New York).
Bernheim, B.D. (1997c), "Rethinking saving incentives", in: A. Auerbach ed., Fiscal Policy: Lessons from Economic Research (MIT Press, Cambridge, MA) pp. 259–311.
Bernheim, B.D. (1998), "Financial illiteracy, education, and retirement saving", in: Olivia S. Mitchell and Sylvester J. Schieber, eds., Living with Defined Contribution Pensions (University of Pennsylvania Press, Philadelphia) pp. 38–68.
Bernheim, B.D., and K. Bagwell (1988), "Is everything neutral?" Journal of Political Economy 96(2):308–338.
Bernheim, B.D., and D.M. Garrett (2002), "The effects of financial education in the workplace: evidence from a survey of households", Journal of Public Economics, forthcoming.
Bernheim, B.D., and J.K. Scholz (1993a), "Private pensions and household saving", Mimeo (University of Wisconsin).
Bernheim, B.D., and J.K. Scholz (1993b), "Private saving and public policy", Tax Policy and the Economy 7:73–110.
Bernheim, B.D., and S. Severinov (2000), "Bequests as signals: an explanation for the equal division puzzle", Working Paper 7791 (National Bureau of Economic Research).
Bernheim, B.D., A. Shleifer and L.H. Summers (1985), "The strategic bequest motive", Journal of Political Economy 93(6):1045–1076.
Bernheim, B.D., J. Skinner and S. Weinberg (2001), "What accounts for the variation in retirement wealth among U.S. households?", American Economic Review 91(4):832–857.
Blinder, A.S. (1974), Toward an Economic Theory of Income Distribution (MIT Press, Cambridge, MA).
Blinder, A.S. (1975), "Distribution effects and the aggregate consumption function", Journal of Political Economy 83:447–475.
Blinder, A.S., R.H. Gordon and D.E. Wise (1980), "Reconsidering the work disincentive effects of social security", National Tax Journal 33(4):431–442.
Bloom, D.E., and R.B. Freeman (1992), "The fall in private pension coverage in the U.S.", Working Paper 3973 (National Bureau of Economic Research).
Börsch-Supan, A. (1994), "Savings in Germany – Part 1: incentives", in: James M. Poterba, ed., Public Policies and Household Saving (University of Chicago Press, Chicago) pp. 81–104.
Boskin, M. (1978), "Taxation, saving, and the rate of interest", Journal of Political Economy 86:S3–S27.
Boskin, M., and L.J. Lau (1978), "Taxation, social security and aggregate factor supply in the United States", Mimeo (Stanford University).
Buck Consultants, Inc. (1989), Current 401(k) Plan Practices: A Survey Report (Buck Consultants Inc., New York).

Bull, N. (1993), "When all the optimal dynamic taxes are zero", Mimeo (Board of Governors of the Federal Reserve System, Washington, D.C.)

Burbidge, J.B., and J.B. Davies (1994), "Government incentives and household saving in Canada", in: James M. Poterba, ed., Public Policies and Household Saving (University of Chicago Press, Chicago) pp. 1–56.

Burbidge, J.B., D. Fretz and M.R. Veall (1998), "Canadian and American saving rates and the role of RRSPs", Canadian Public Policy 24(2):259–263.

Burkhauser, R.V. (1979), "The pension acceptance of older workers", Journal of Human Resources 14:63–72.

Burkhauser, R.V. (1980), "The early acceptance of social security: an asset maximization approach", Industrial and Labor Relations Review 33:484–492.

Burman, L., J. Cordes and L. Ozanne (1990), "IRAs and national saving", National Tax Journal 43:259–284.

Cagan, P. (1965), "The effect of pension plans on aggregate saving: evidence from a sample survey", Occasional Paper 95 (National Bureau of Economic Research).

Campbell, J.Y., and N.G. Mankiw (1989), "Consumption, income and interest rates: reinterpreting the time series evidence", in: Olivier Jean Blanchard and Stanley Fischer, eds., NBER Macroeconomics Annual (MIT Press, Cambridge, MA) pp. 185–216.

Carroll, C., and L.H. Summers (1987), "Why have private savings rates in the U.S. and Canada diverged"? Journal of Monetary Economics 20:249–280.

Central Council for Savings Promotion (1981), Savings and Savings Promotion Movement in Japan (Bank of Japan, Tokyo, Japan).

Chamley, C. (1981), "The welfare cost of capital income taxation in a growing economy", Journal of Political Economy 89:468–496.

Chamley, C. (1986), "Optimal taxation of capital income in general equilibrium with infinite lives", Econometrica 54(3):607–622.

Chari, V.V., and P.J. Kehoe (1999), "Optimal fiscal and monetary policy", in: J.B. Taylor and M. Woodford, eds., Handbook of Macroeconomics (Elsevier, Amsterdam) Chapter 26.

Chari, V.V., L.J. Christiano and P.J. Kehoe (1994), "Optimal fiscal policy in a business cycle model", Journal of Political Economy 102(4):617–652.

Clark, R.L., and A. McDermed (1990), The Choice of Pension Plans in a Changing Regulatory Environment (American Enterprise Institute, Washington, D.C).

Clark, R.L., and S.J. Schieber (1998), "Factors affecting participation rates and contribution levels in 401(k) plans", in: Olivia S. Mitchell and Sylvester J. Schieber, eds., Living with Defined Contribution Pensions (University of Pennsylvania Press, Philadelphia) pp. 69–97.

Conlisk, J. (1996), "Why bounded rationality"? Journal of Economic Literature 34(2):669–700.

Cox, D. (1987), "Motives for private income transfers", Journal of Political Economy 95:508–546.

Davies, J.B. (1981), "Uncertain lifetime, consumption, and dissaving in retirement", Journal of Political Economy 89:561–577.

Deaton, A. (1992), Understanding Consumption (Clarendon Press, Oxford, England).

Diamond, P.A. (1970), "Incidence of an Interest Income Tax", Journal of Economic Theory 2(3):211–224.

Diamond, P.A. (1973), "Taxation and public production in a growth setting", in: J.A. Mirrlees and N.H. Stern, eds., Models of Economic Growth (Macmillan, London) pp. 215–234.

Diamond, P.A., and J.A. Hausman (1984), "Individual retirement and saving behavior", Journal of Public Economics 23:81–114.

Dicks-Mireaux, L., and M.A. King (1984), "Pension wealth and household savings: tests of robustness", Journal of Public Economics 23:115–139.

Doyle Jr, R.J., and E.T. Johnson (1991), Readings in Wealth Accumulation Planning, fourth Edition (The American College, Bryn Mawr, PA).

Duflo, E., and E. Saez (2000), "Participation and investment decisions in a retirement plan: the influence of colleagues' choices", Working Paper 7735 (National Bureau of Economic Research).
Dynan, K.E. (1993), "How prudent are consumers?" Journal of Political Economy 101(6):1104–1113.
Elmendorf, D.W. (1996), "The effect of interest-rate changes on household saving and consumption: a survey", Mimeo (Federal Reserve Board, Washinton, D.C.)
Employee Benefit Research Institute (1995), "Can we save enough to retire? Participant education in defined contribution plans", EBRI Issue Brief 160 (Employee Benefit Research Institute, Washington, D.C.)
Engelhardt, G.V. (1996), "Tax subsidies and household saving: evidence from Canada", Quarterly Journal of Economics 3(4):1237–1268.
Engen, E.M. (1994), "Precautionary saving and the structure of taxation", Mimeo (Federal Reserve Board of Governors, Washington, D.C.)
Engen, E.M., and W.G. Gale (1993), "IRAs and saving in a stochastic life-cycle model", Mimeo (Brookings Institution, Washington, D.C.)
Engen, E.M., and W.G. Gale (1996a), "Taxation and saving: the role of uncertainty", Mimeo (Federal Reserve Board of Governors, Washington, D.C.)
Engen, E.M., and W.G. Gale (1996b), "Comprehensive tax reform and the private pension system", Mimeo (Federal Reserve Board, Washington, D.C.)
Engen, E.M., and W.G. Gale (1997), "Debt, taxes, and the effects of 401(k) plans on household wealth accumulation", Mimeo (Federal Reserve Board, Washington, D.C.)
Engen, E.M., W.G. Gale and J.K. Scholz (1994), "Do saving incentives work"? Brookings Papers on Economic Activity 1:85–151.
Engen, E.M., W.G. Gale and J.K. Scholz (1996a), "The illusory effects of saving incentives on saving", Journal of Economic Perspectives 10(4):113–138.
Engen, E.M., W.G. Gale and J.K. Scholz (1996b), "Effects of tax-based saving incentives on saving and wealth: a critical review of the literature", Working Paper 5759 (National Bureau of Economic Research).
Evans, O.J. (1983), "Tax policy, the interest elasticity of saving, and capital accumulation: numerical analysis of theoretical models", American Economic Review 74:398–409.
Feenberg, D.R., and J. Skinner (1989), "Sources of IRA saving", Tax Policy and the Economy 3:25–46.
Feldstein, M. (1973), "Tax incentives, corporate saving, and capital accumulation in the United States", Journal of Public Economics 2:159–171.
Feldstein, M. (1978), "The welfare cost of capital income taxation", Journal of Political Economy 86(2):S29–51.
Feldstein, M. (1994), "The effect of marginal tax rates on taxable income: a panel study of the 1986 Tax Reform Act", Working Paper 4496 (National Bureau of Economic Research).
Feldstein, M., and G. Fane (1973), "Taxes, corporate dividend policy and personal savings: the British postwar experience", Review of Economics and Statistics 55(4):399–411.
Fields, G.S., and O.S. Mitchell (1984), Retirement, Pensions, and Social Security (MIT Press, Cambridge, MA).
Fisher, I. (1930), The Theory of Interest (MacMillan, London).
Fougère, D. (1994), "Public policies and household saving in France", in: James M. Poterba, ed., Public Policies and Household Saving (University of Chicago Press, Chicago) pp. 161–190.
Fullerton, D., and D.L. Rogers (1993), Who Bears the Lifetime Tax Burden? (Brookings Institution, Washington, D.C.).
Fullerton, D., and D.L. Rogers (1996), "Lifetime effects of fundamental tax reform", in: Henry J. Aaron and William G. Gale, eds., Economic Effects of Fundamental Tax Reform (Brookings Institution, Washington, D.C.) pp. 321–347.
Furnham, A., and A. Lewis (1986), The Economic Mind: The Social Psychology of Economic Behavior (St. Martin's Press, New York).

Gale, W.G. (1995), "The effects of pensions on wealth: a re-evaluation of theory and evidence", Mimeo (Brookings Institution).
Gale, W.G., and J.K. Scholz (1994a), "Intergenerational transfers and the accumulation of wealth", Journal of Economic Perspectives 8:145–160.
Gale, W.G., and J.K. Scholz (1994b), "IRAs and household saving", American Economic Review 84:1233–1260.
Garrett, D.M. (1995), "The effects of nondiscrimination rules on 401(k) contributions", Mimeo (Stanford University).
Gentry, W., and E. Peress (1994), "Taxes and fringe benefits offered by employers", Working Paper 4764 (National Bureau of Economic Research).
Gravelle, J.G. (1991a), "Income, consumption, and wage taxation in a life cycle model: separating efficiency from redistribution", American Economic Review 81:985–995.
Gravelle, J.G. (1991b), "Do individual retirement accounts increase savings"? Journal of Economic Perspectives 5:133–148.
Green, F. (1981), "The effect of occupational pension schemes on saving in the united kingdom: a test of the life cycle hypothesis", The Economic Journal 91:136–144.
Gustman, A.L., and T.L. Steinmeier (1992), "The stampede toward defined contribution pension plans: fact or fiction"? Industrial Relations 31:361–369.
Gustman, A.L., and T.L. Steinmeier (1995), Pension Incentives and Job Mobility (W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan).
Gustman, A.L., and T.L. Steinmeier (1998), "Effects of pensions on savings: analysis with data from the health and retirement survey", Working Paper 6681 (National Bureau of Economic Research).
Hall, R.E. (1988), "Intertemporal substitution in consumption", Journal of Political Economy 96: 339–357.
Harberger, A.C. (1964), "Taxation, resource allocation and welfare", in: The Role of Direct and Indirect Taxes in the Federal Revenue System (Princeton University Press, Princeton, NJ) pp. 25–70.
Hayashi, F. (1985), "Tests for liquidity constraints: a critical survey", Working Paper 1720 (National Bureau of Economic Research).
Hoch, S.J., and G.F. Lowenstein (1991), "Time-inconsistent preferences and consumer self-control", Journal of Consumer Research 17(4):492–507.
Howrey, E.P., and S.H. Hymans (1978), "The measurement and determination of loanable funds saving", Brookings Papers on Economic Activity 2:655–685.
Hubbard, R.G. (1984), "Do IRAs and Keoghs increase saving"? National Tax Journal 37:43–54.
Hubbard, R.G. (1986), "Pension wealth and individual saving", Journal of Money, Credit, and Banking 18(2):167–178.
Hubbard, R.G., and K.L. Judd (1986), "Liquidity constraints, fiscal policy, and consumption", Brookings Papers on Economic Activity 1:1–51.
Hubbard, R.G., and J. Skinner (1996), "Assessing the effectiveness of saving incentives", Journal of Economic Perspectives 10(4):73–90.
Hurd, M. (1987), "Saving of the elderly and desired bequests", American Economic Review 77:298–312.
Hurd, M. (1989), "Mortality risk and bequests", Econometrica 57:779–814.
Ippolito, R.A. (1985), "The economic function of underfunded pension plans", Journal of Law and Economics 28:611–651.
Ippolito, R.A. (1986), Pensions, Economics and Public Policy (Dow Jones-Irwin, Homewood, IL).
Ito, T., and Y. Kitamura (1994), "Public policies and household saving in Japan", in: James M. Poterba, ed., Public Policies and Household Saving (University of Chicago Press, Chicago) pp. 133–160.
Jappelli, T., and M. Pagano (1994), "Government incentives and household saving in Italy", in: James M. Poterba, ed., Public Policies and Household Saving (University of Chicago Press, Chicago) pp. 105–132.
Johnson, A.P. (1985), "Individual retirement accounts help boost saving in the U.S.", Testimony to

the Committee on Finance, U.S. Senate, in: Tax Reform Proposals 99–246(part XIII) (US GPO, Washington, D.C.) pp. 129–149.
Johnson, C.E., J. Diamond and G.R. Zodrow (1997), "Bequests, saving, and taxation", in: Proceedings of the Eighty-Ninth Annual Conference on Taxation, 1996 (National Tax Association, Washington, D.C.) pp. 37–45.
Joines, D.H., and J.G. Manegold (1995), "IRA and saving: evidence from a panel of taxpayers", Mimeo (University of Southern California).
Jones, L.E., R.E. Manuelli and P.E. Rossi (1993), "Optimal taxation in models of endogenous growth", Journal of Political Economy 101(3):485–517.
Jones, L.E., R.E. Manuelli and P.E. Rossi (1997), "On the optimal taxation of capital income", Journal of Economic Theory 73:93–117.
Jones, S.R.G. (1984), The Economics of Conformism (Basil Blackwell, Oxford).
Judd, K.L. (1985), "Redistributive taxation in a simple perfect foresight model", Journal of Public Economics 28:59–83.
Judd, K.L. (1987), "The welfare cost of factor taxation in a perfect-foresight model", Journal of Political Economy 95(4):675–709.
Judd, K.L. (1997), "The optimal tax rate for capital income is negative", Working Paper 6004 (National Bureau of Economic Research).
Judd, K.L. (1999), "Optimal taxation and spending in general competitive growth models", Journal of Public Economics 71(1):1–25.
Katona, G. (1965), Private Pensions and Individual Saving (University of Michigan Press, Ann Arbor, MI).
Katona, G. (1975), Psychological Economics (Elsevier, Amsterdam).
King, M.A., and L. Dicks-Mireaux (1982), "Asset holdings and the life cycle", Economic Journal 92:247–267.
Kotlikoff, L.J. (1979), "Testing the theory of social security and life cycle accumulation", American Economic Review 69(3):396–410.
Kotlikoff, L.J. (1988), "Intergenerational transfers and savings", Journal of Economic Perspectives 2:41–58.
Kotlikoff, L.J., and A. Spivak (1981), "The family as an incomplete annuities market", Journal of Political Economy 89:372–391.
Kotlikoff, L.J., and L.H. Summers (1981), "The role of intergenerational transfers in aggregate capital accumulation", Journal of Political Economy 89:706–732.
Kotlikoff, L.J., and D.A. Wise (1989), The Wage Carrot and the Pension Stick: Retirement Benefits and Labor Force Participation (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).
Kruse, D.L. (1995), "Pension substitution in the 1980s: why the shift toward defined contribution pension plans"? Industrial Relations 34(2):218–241.
Kurz, M. (1984), "Capital accumulation and the characteristics of private intergenerational transfers", Economica 51:1–22.
Kusko, A., J.M. Poterba and D.W. Wilcox (1998), "Employee decisions with respect to 401(k) plans", in: O.S. Mitchell and S.J. Schieber, eds., Living with Defined Contribution Pensions (University of Pennsylvania Press, Phildelphia) pp. 98–112.
Laibson, D.I. (1994a), "Self-control and saving", Mimeo (Harvard University).
Laibson, D.I. (1994b), "Mental accounts, self-control and an intrapersonal principal-agent problem", Mimeo (Harvard University).
Laibson, D.I. (1996), "Hyperbolic discount functions, undersaving and savings policy", Working Paper 5635 (National Bureau of Economic Research).
Laibson, D.I. (1998), "Comment on personal retirement saving programs and asset accumulation", in: David A. Wise, ed., Studies in the Economics of Aging (NBER, Cambridge, MA, and the University of Chicago Press, Chicago) pp. 106–124.
Laibson, D.I., A. Repetto and J. Tobacman (1998), "Self-control and retirement savings", Brookings Papers on Economic Activity 1:91–196.

Laitner, J., and F.T. Juster (1996), "New evidence on altruism: a study of TIAA-CREF retirees", American Economic Review 86:893–908.
Lawrance, E.C. (1991), "Poverty and the rate of time preference: evidence from panel data", Journal of Political Economy 99(1):54–77.
Lazear, E.P. (1984), "Pensions as severance pay", in: Z. Bodie, J. Shoven and D. Wise, eds., Financial Aspects of the United States Pension System (University of Chicago Press, Chicago) pp. 57–85.
Lazear, E.P., and R. Moore (1988), "Pensions and turnover", in: Z. Bodie, J. Shoven and D. Wise, eds., Pensions in the U.S. Economy (University of Chicago Press, Chicago) pp. 163–190.
Lipman, B.L. (1991), "How to decide how to decide how to... modeling limited rationality", Econometrica 59(4):1105–1125.
Long, J.E. (1990), "Marginal tax rates and IRA contributions", National Tax Journal 43(2):143–153.
Lord, W., and P. Rangazas (1992), "Tax reform with altruistic bequests", Public Finance 47:61–81.
Lucas Jr, R.E. (1976), "Econometric policy evaluation: a critique", in: Karl Brunner and Allan H. Meltzer, eds., The Phillips Curve and Labor Markets, Vol. 1, Carnegie-Rochester Conference Series on Public Policy, Journal of Monetary Economics (suppl.).
Maital, S. (1986), "Prometheus rebound: on welfare-improving constraints", Eastern Economic Journal 12(3):337–344.
McGee, M.K. (1989), "Alternative transitions to a consumption tax", National Tax Journal 42:155–166.
Menchik, P.L. (1980), "Primogeniture, equal sharing and the U.S. distribution of wealth", Quarterly Journal of Economics 94:299–316.
Milligan, K.S. (1998), "Savings and tax incentives: semiparametric estimation of the household savings impact of RRSPs", Mimeo (University of Toronto).
Modigliani, F. (1988), "The role of intergenerational transfers and life cycle saving in the accumulation of wealth", Journal of Economic Perspectives 2:15–40.
Modigliani, F., and R. Brumberg (1954), "Utility analysis and the consumption function: an interpretation of cross-section data", in: K.K. Kurihara, ed., Post Keynesian Economics (Rutgers University Press, New Brunswick, NJ) pp. 388–436.
Munnell, A.H. (1974), The Effect of Social Security on Personal Saving (Ballinger, Cambridge, MA).
Munnell, A.H. (1976), "Private pensions and saving: new evidence", Journal of Political Economy 84(5):1013–1032.
Papke, L. (1992), "Participation in and contributions to 401(k) plans: evidence from plan data", Working Paper 4199 (National Bureau of Economic Research).
Papke, L., M. Petersen and J.M. Poterba (1993), "Did 401(k) plans replace other employer provided pensions?" Working Paper 4501 (National Bureau of Economic Research).
Parsons, D.O. (1986), "The employment relationship: job attachment, work effort, and the nature of contracts", in: O. Ashenfelter and R. Layard, eds., Handbook of Labor Economics, Vol. II (North Holland, Amsterdam) pp. 789–848.
Parsons, D.O. (1995), "Retirement age and retirement income: the role of the firm", Mimeo (Ohio State University).
Phelps, E.S., and R.A. Pollak (1968), "On second-best national saving and game-equilibrium growth", Review of Economic Studies 35:185–199.
Poterba, J.M. (1987), "Tax policy and corporate saving", Brookings Papers on Economic Activity 2:455–503.
Poterba, J.M. (1991), "Dividends, capital gains, and the corporate veil: evidence from Britain, Canada, and the United States", in: B. Douglas Bernheim and John B. Shoven, eds., National Saving and Economic Performance (University of Chicago Press, Chicago) pp. 49–74.
Poterba, J.M., S.F. Venti and D.A. Wise (1992), "401(k) plans and tax-deferred saving", Working Paper 4181 (National Bureau of Economic Research).
Poterba, J.M., S.F. Venti and D.A. Wise (1994), "401(k) plans and tax-deferred saving", in: D.A. Wise, ed., Studies in the Economics of Aging (University of Chicago Press, Chicago) pp. 105–138.

Poterba, J.M., S.F. Venti and D.A. Wise (1995), "Do 401(k) contributions crowd out other personal saving?" Journal of Public Economics 58:1–32.
Poterba, J.M., S.F. Venti and D.A. Wise (1996a), "How retirement saving programs increase saving", Journal of Economic Perspectives 10(4):91–112.
Poterba, J.M., S.F. Venti and D.A. Wise (1996b), "Personal retirement saving programs and asset accumulation: reconciling the evidence", Working Paper 5599 (National Bureau of Economic Research).
Quinn, J.F., R.V. Burkhauser and D.A. Myers (1990), Passing the Torch: The Influence of Economic Incentives on Work and Retirement (W.E. Upjohn Institute, Kalamazoo, MI).
Rainwater, L. (1970), Behind Ghetto Walls: Black Families in a Federal Slum (Aldine, Chicago).
Reagan, P.B., and J.A. Turner (1995), "The decline in marginal tax rates during the 1980s reduced pension coverage", Mimeo (Ohio State University).
Runkle, D.E. (1991), "Liquidity constraints and the permanent-income hypothesis: evidence from panel data", Journal of Monetary Economics 27(1):73–98.
Sabelhaus, J. (1997), "Public policy and saving in the United States and Canada", Canadian Journal of Economics 30(2):253–275.
Samwick, A. (1995), "The limited offset between pension wealth and other private wealth: implications of buffer-stock saving", Mimeo (Dartmouth College, Hanover, NH).
Sandmo, A. (1985), "The effects of taxation on savings and risk taking", in: A.J. Auerbach and M. Feldstein, eds., Handbook of Public Economics, Vol. 1 (North Holland, Amsterdam) pp. 265–311.
Schelling, T.C. (1984), "Self-command in practice, in policy, and in a theory of rational choice", American Economic Review 74(2):1–11.
Schoeplein, R.N. (1970), "The effect of pension plans on other retirement saving", Journal of Finance 25:633–638.
Scitovsky, T. (1976), The Joyless Economy (Oxford University Press, Oxford).
Scott, J. (1994), "The determinants of participation in defined contribution pension plans", Mimeo (Stanford University).
Seidman, L.S. (1983), "Taxes in a life cycle growth model with bequests and inheritances", American Economic Review 93:437–441.
Seidman, L.S. (1984), "Conversion to a consumption tax: the transition in a life-cycle growth model", Journal of Political Economy 92:247–267.
Shapiro, M.D. (1984), "The permanent income hypotheses and the real interest rate: some evidence from panel data", Economic Letters 14:93–100.
Shefrin, H., and R.H. Thaler (1988), "The behavioral life-cycle hypothesis", Economic Inquiry 26(4):609–643.
Shoven, J.B. (1991), Return on Investment: Pensions are How America Saves (Association of Private Pension and Welfare Plans, Washington D.C.)
Simon, H.A. (1955), "A behavioral model of rational choice", Quarterly Journal of Economics 69(1):99–118.
Skinner, J., and D.R. Feenberg (1989), "The impact of the 1986 Tax Reform on personal saving", in: Joel Slemrod, ed., Do Taxes Matter? The Effect of the 1986 Tax Reform Act on the U.S. Economy (MIT Press, Cambridge).
Stack, C.B. (1974), All Our Kin: Strategies for Survival in a Black Community (Harper and Row, New York).
Starrett, D.A. (1988), "Effects of taxes on saving", in: H. Galper, H. Aaron and J. Pechman, eds., Uneasy Compromise: Problems of a Hybrid Income–Consumption Tax (The Brookings Institution, Washington, D.C.) pp. 237–259.
Stock, J.H., and D.A. Wise (1990), "The pension inducement to retire: an option value analysis", in: D.A. Wise, ed., Issues in the Economics of Aging (University of Chicago Press, Chicago) pp. 205–224.
Strotz, R.H. (1955), "Myopia and inconsistency in dynamic utility maximization", Review of Economic Studies 23:165–180.

Summers, L.H. (1981), "Taxation and capital accumulation in a life cycle growth model", American Economic Review 71:533–554.
Summers, L.H. (1986), "Summers replies to Galper and Byce on IRAs", Tax Notes 31(10):1014–1016.
Thaler, R.H. (1994), "Psychology and savings policies", American Economic Review 84:186–192.
Thaler, R.H., and H.M. Shefrin (1981), "An economic theory of self-control", Journal of Political Economy 89(2):392–406.
Tomes, N. (1981), "The family, inheritance, and the intergenerational transmission of inequality", Journal of Political Economy 89:928–958.
US Congressional Budget Office (1993), Baby Boomers in Retirement: An Early Perspective (Government Printing Office, Washington, D.C.)
Venti, S.F., and D.A. Wise (1986), "Tax-deferred accounts, constrained choice, and estimation of individual saving", Review of Economic Studies 53:579–601.
Venti, S.F., and D.A. Wise (1987), "IRAs and saving", in: Martin Feldstein, ed., The Effects of Taxation on Capital Accumulation (University of Chicago Press, Chicago).
Venti, S.F., and D.A. Wise (1988), "The determinants of IRA contributions and the effect of limit changes", in: Z. Bodie, J.B. Shoven and D.A. Wise, eds., Pensions and the U.S. Economy (University of Chicago Press, Chicago) pp. 9–47.
Venti, S.F., and D.A. Wise (1990), "Have IRAs increased U.S. saving: evidence from consumer expenditure surveys", Quarterly Journal of Economics 105:661–698.
Venti, S.F., and D.A. Wise (1991), "The saving effect of tax-deferred retirement accounts: evidence for SIPP", in: B.D. Bernheim and J.B. Shoven, eds., National Saving and Economic Performance (University of Chicago Press, Chicago) pp. 103–128.
Venti, S.F., and D.A. Wise (1992), "Government policy and personal retirement saving", Tax Policy and the Economy 6:1–41.
Venti, S.F., and D.A. Wise (1993), "The wealth of cohorts: retirement saving and the changing assets of older Americans", Working Paper 4600 (National Bureau of Economic Research).
Venti, S.F., and D.A. Wise (1995a), "Individual response to retirement savings programs: results from U.S. panel data", Reicherche Economiche 49(3):235–254.
Venti, S.F., and D.A. Wise (1995b), "RRSPs and saving in Canada", Mimeo (Dartmouth College, Hanover, NH).
Whyte, W.F. (1943), Street Corner Society (University of Chicago Press, Chicago).
Wilhelm, M.O. (1996), "Bequest behavior and the effect of heirs' earnings: testing the altruistic model of bequests", American Economic Review 86:874–892.
Williamson, S.H. (1992), "U.S. and Canadian pensions before 1930: a historical perspective", in: J.A. Turner and D.J. Beller, eds., Trends in Pensions 1992 (U.S. Department of Labor, Pension and Welfare Benefits Administration, Washington, D.C.) pp. 35–57.
Wolff, E.N. (1988), "Social security, pensions and the life cycle accumulation of wealth: some empirical tests", Annales d'Economie et de Statistique 9:199–226.
Woodbury, S.A., and D.R. Bettinger (1991), "The decline of fringe-benefit coverage in the 1980s", in: Randall W. Eberts and Erica L. Groshen, eds., Structural Changes in the U.S. Labor Markets: Causes and Consequences (M.E. Sharpe, Armonk, New York) pp. 105–138.
Woodbury, S.A., and W.-J. Huang (1993), The Tax Treatment of Fringe Benefits (Upjohn Institute, Kalamazoo, Michigan).
Wright, C. (1969), "Saving and the rate of interest", in: A.C. Harberger and M.J. Bailey, eds., The Taxation of Income from Capital (Brookings Institution, Washington D.C.) pp. 275–300.
Yotsuzuka, T. (1987), "Ricardian equivalence in the presence of capital market imperfections", Journal of Monetary Economics 20(2):411–436.
Zeldes, S.P. (1989), "Consumption and liquidity constraints: an empirical investigation", Journal of Political Economy 94(2):305–346.
Zhu, X. (1992), "Optimal fiscal policy in a stochastic growth model", Journal of Economic Theory 58(2):250–289.

Chapter 19

TAXATION AND CORPORATE FINANCIAL POLICY*

ALAN J. AUERBACH

University of California, Berkeley and NBER

Contents

* I am grateful to Kevin Cole for research assistance, to the Burch Center for Tax Policy and Public Finance for research support, and to Doug Bernheim, John Graham, Jim Hines, Vesa Kanniainen, Hans-Werner Sinn and Jan Södersten for comments on an earlier draft.

Handbook of Public Economics, Volume 3, Edited by A.J. Auerbach and M. Feldstein

Abstract

This chapter reviews the theory and evidence regarding the impact of taxation on corporate financial policy. Starting from a basic characterization of the classical corporate income tax and its effects, the analysis focuses on three areas of research: equity policy, debt–equity decisions, and choices regarding ownership structure and organizational form. The discussion stresses the distinction between nominal and more fundamental financial differences – for example, in the relationship between borrowing and leasing – and that financial policy involves choices not only among different underlying policies but also among characterizations of a given policy. The final section offers some brief reflections on the implications of continuing financial innovation.

Keywords

Modigliani–Miller theorem, debt–equity ratio, leasing, clientele effect, internal funds

JEL classification: H32, G3

1. Introduction

Like other countries that rely on the income tax as a source of revenue, the United States distinguishes between corporations and individuals. US corporations and individuals face separate tax schedules and different rules regarding income and deductions. Under this *classical* system of corporate taxation, there is limited coordination, or *integration,* of the two tax systems: taxes on shareholders are assessed independently of the taxes on the corporations they own. By contrast, many other countries have attempted to effect some form of integration of corporate and individual income taxes. However, even in these countries, adjustments have taken the form of partial measures, leaving the corporation income tax with independent effects.

Through the years, economists have devoted considerable effort to understanding the incidence of a distinct corporation income tax and its impact on the investment and financial decisions of firms. This chapter reviews the portion of this literature that has focused on corporate financial policy, including choices about firm ownership structure. Other chapters in this Handbook, by Fullerton and Metcalf (Volume 4, forthcoming) and by Hassett and Hubbard (20), consider more fully the issues of incidence and investment, respectively. Poterba's chapter (17) focuses on the effects of taxation on the financial decisions of households, rather than firms, and Gordon and Hines (Volume 4, forthcoming) deal with the considerable complications introduced by open-economy capital movements. However, the discussion below must, of necessity, touch on the issues covered more fully in these other chapters. The incidence of the corporation income tax depends, in part, on the nature of financial equilibrium; the real and financial decisions of firms are independent only under restrictive assumptions; corporate financial decisions should be sensitive to the taxes faced by the owners and potential owners of their securities; and the domestic financial equilibrium will depend on tax rules that influence foreign capital flows.

1.1. What is financial policy?

In the simplest terms, financial policy relates to two key choices that firms make: (1) how much of their capital structure to support by debt, rather than equity; and (2) how much of their earnings to retain for use as internal equity finance, rather than distributing dividends and raising new equity in the market. In two landmark papers, Modigliani and Miller (1958) and Miller and Modigliani (1961) demonstrated, under certain assumptions, that neither of these decisions mattered, having no effect on firm value and shareholder wealth. These papers launched the modern literature on corporate financial policy, establishing a benchmark against which deviations from the M–M assumptions – such as the existence of taxes – could be evaluated.

The key insight of the M–M analysis is that market valuations should relate to underlying claims to income streams, rather than to how assets are labeled. A portfolio consisting of a little risky equity and a lot of safe debt should have the same value as a second portfolio with a lot of less risky equity and a little safe debt if the underlying

risk of the two portfolios is comparable. We should go beyond terms like "debt" and "equity" to consider the characteristics of the claims themselves.

Over the years, this lesson has been emphasized by the evolution of financial instruments such as leases, which may act as substitutes for debt, and options, the valuation of which can, once again, be understood by constructing comparable portfolios with and without options and requiring that they have the same value [Black and Scholes (1973)]. A challenge to analyzing the impact of taxation on firm decisions, though, is that the tax system is based in large part on formal labels, and only indirectly on underlying asset characteristics. Thus, equity faces one set of tax rules and debt another, often more favorable, so special rules are needed regarding the treatment of the risky debt that more closely resembles equity. Equity repurchases are treated more favorably than are dividends but, again, restrictions exclude from this favorable treatment share redemptions that too closely resemble dividends.

Evaluating the impact of taxes on firm behavior requires that we understand the rules that apply in distinguishing among different types of assets. Financial policy decisions often amount to choosing the optimal trade-off between distortions to financial policy and the tax benefits such distortions generate. Indeed, a major tax-avoidance activity consists of trying to improve this trade-off, constructing assets and transactions to permit corporations to characterize their financial decisions in a manner most favorable from the tax standpoint. The impact of taxation, then, depends not only on the tax system itself, but also on where the tax system's definitional lines are drawn and how well they can be "moved" through tax-avoidance activity.

1.2. Outline of the chapter

Each of the three sections that follow deals with an important aspect of corporate financial policy, respectively equity policy, debt–equity decisions, and choices regarding ownership structure and organizational form. The final section offers some brief reflections on the implications of continuing financial innovation. The discussion below relies heavily on my previous survey paper [Auerbach (1983a)] with respect to developments in the literature up to that paper's writing, and on the section in Auerbach and Slemrod (1997) concerning the impact of the Tax Reform Act of 1986 on financial policy.

2. Corporate equity policy

While risk is an essential component of the theory of corporate financial decisions, a useful starting place to analyze the effects of taxation is a model without risk. Also eschewing for the moment the important question of investor heterogeneity, we consider the behavior of a representative firm whose securities are owned by a representative individual, with firm and individual each facing its own, distinct tax system, and no provisions that integrate the two. The basic approach follows that

laid out in King (1974, 1977), Auerbach (1979b), Edwards and Keen (1984) and, for the continuous-time analogue used here, Sinn (1987).

Corporations face a single income tax rate, τ, which will enter the analysis later, while individuals face distinct tax rates θ on dividends and c ($\leqslant \theta$) on accrued capital gains. In reality, capital gains are generally taxed on realization rather than on accrual, a distinction that is important from the perspective of household portfolio reallocation decisions. However, incorporating a realization-based capital-gains tax would complicate the present analysis greatly, and is not as important in this context. The accrual-equivalent alternative, c, should be thought of as being considerably less than the actual capital-gains tax rate, because it takes into account the fact that not all gains are realized in every year, and that gains realized in the future benefit from a deferral advantage[1].

Let V_t be the value of the firm at time t. It is also useful to introduce the measure S_t to represent the value of new shares issued at date t. If $S_t < 0$, then the firm is a net repurchaser of its own shares. Let D_t be the firm's total dividend payment at date t, and let ρ be the discount rate that the representative investor applies to the cash flows and capital gains generated by the firm. Capital-market equilibrium requires that the after-tax rate of return equal ρ:

$$\rho = \frac{D_t}{V_t}(1-\theta) + \frac{\dot{V}_t - S_t}{V_t}(1-c), \tag{2.1}$$

where $\dot{V}_t$ is the rate of change of V_t with respect to time, t, and where the second term on the right-hand side of Equation (2.1) reflects the fact that increases in share values due to *extensive* growth through share issuance are not taxable.

Rewriting Equation (2.1) as a simple first-order differential equation in V,

$$\frac{\rho}{1-c}V_t = \dot{V}_t + D_t\left(\frac{1-\theta}{1-c}\right) - S_t, \tag{2.2}$$

and solving forward using the terminal condition that discounted firm value converge to zero, we obtain the following expression for firm value at date t:

$$V_t = \int_t^\infty \exp\left(-\frac{\rho}{1-c}(s-t)\right)\left[D_s\left(\frac{1-\theta}{1-c}\right) - S_s\right] ds. \tag{2.3}$$

Expression (2.3) is valid for any path of dividends and share issues, and so can serve as a basis for determining the optimal choices of these two variables to maximize firm value. These choices are not independent, and are further constrained

[1] King (1977) discusses the construction of accrual equivalent measures. Poterba's chapter in this Handbook discusses capital gains taxes and their effects.

by technological and legal constraints on the firm. The most obvious constraint is that imposed by the firm's net cash flow: net cash leaving the firm equals dividends less net new share issues. If we define G_t as the net proceeds from the firm's operations before the determination of dividends and new share issues, then this constraint is

$$G_t \equiv D_t - S_t. \tag{2.4}$$

In addition, dividends cannot be negative ($D_t \geqslant 0$). However, there may be further constraints on the payment of dividends. For example, one might imagine firms finding it necessary to pay out a certain share of their earnings as dividends. As discussed further below, the motivation for such behavior requires a richer model than the current one, notably some combination of asymmetric information and a divergence of interests between shareholder and corporate manager. However, for the moment, we can simply consider the implications of imposing such a constraint, as in:

$$D_t \geqslant p(D_t + \dot{V}_t - S_t), \tag{2.5}$$

which requires that dividends equal at least a fraction p of the firm's total returns[2].

There may also be effective restrictions on share repurchases, which have the attraction over dividends of facing capital-gains tax rates. Although there have been legal restrictions on repurchases elsewhere, impediments in the United States are limited to taxation, treating repurchases as dividends if they are distributed in proportion to share ownership. While other methods of repurchasing (via the open market or through tender offers) are unlikely to result in proportional sales by different investors, repurchases have, except during certain periods, been uncommon relative to new share issues and dividends. This suggests that there may be factors beyond those explicit in the model that limit a firm's ability to repurchase its shares, by making it costly to do so.

Perhaps most importantly, repurchases from investors who voluntarily tender their shares may also be subject to the non-tax costs associated with asymmetric information. When firms have the potential to take advantage of tendering shareholders, and an incentive to do so (perhaps in the interest of remaining shareholders) their decision to repurchase equity may attach a premium to the shares they seek to acquire. Barclay and Smith (1988) provide empirical evidence in support of this claim. As suggested by Brennan and Thakor (1990), these costs can lead to a situation in which firms use repurchases only for large distributions, when the advantages of a repurchase overcome the costs of acquiring information about the true value of the firm. As argued by

[2] While this is a particularly simple constraint, imposing a more general cost relating to the dividend payout ratio leads to a similar outcome. See Poterba and Summers (1985). The key is that an increase in earnings leads to an increase in dividends. The same conclusion also applies to the constraint on share repurchases given below in Equation (2.6); the results when that constraint binds are similar to those derived from a more general cost of entering the external equity market.

Myers and Majluf (1984), such costs may be associated with entering the external equity market and hence applicable to new share issues as well, causing share prices to fall upon the announcement of a new issue [Asquith and Mullins (1986)]. But firms impelled to issue new shares have no other source of external equity funds, while those contemplating a repurchase do have the option of paying dividends. Thus, in a richer model in which utilizing the external equity market is costly, we might observe firms issuing equity but not repurchasing equity. We return to this question below but, again, begin simply by considering the impact of such an effective constraint,

$$S_t \geqslant 0. \tag{2.6}$$

To consider the policy that maximizes firm value (2.3) subject to the constraints (2.4)–(2.6), we use Equation (2.4) to substitute for S_t in Equations (2.3), (2.5) and (2.6), and form a Lagrangean:

$$\begin{aligned} V_t = \int_t^\infty & \exp\left(-\frac{\rho}{1-c}(s-t)\right) \\ & \times \left[G_s + D_s\left(\frac{1-\theta}{1-c} - 1\right) + \lambda_s(D_s - p\dot{V}_s - pG_s) + \mu_s(D_s - G_s)\right] \mathrm{d}s, \end{aligned} \tag{2.7}$$

where the multipliers λ_s and μ_s are associated with the constraints (2.5) and (2.6), at least one of which will be binding at any given date.

Expression (2.7) is complicated by the presence of the term $\dot{V}_s$ on the right-hand side. To simplify, we take the derivative of Equation (2.7) with respect to time to obtain a first-order differential equation analogous to Equation (2.2), and solve using the same approach used to reach Equation (2.3). The result is

$$\begin{aligned} V_t = \int_t^\infty & \exp\left(-\int_t^s \frac{\rho}{(1-c)(1-\lambda_v p)}\,\mathrm{d}v\right) \\ & \times \frac{1}{1-\lambda_s p}\left[G_s(1-\lambda_s p - \mu_s) + D_s\left(\frac{1-\theta}{1-c} - 1 + \lambda_s + \mu_s\right)\right] \mathrm{d}s. \end{aligned} \tag{2.8}$$

The first-order condition with respect to dividends, D_s, is:

$$\lambda_s + \mu_s = 1 - \frac{1-\theta}{1-c}, \tag{2.9}$$

so that the second term in brackets on the right-hand side of Equation (2.8) vanishes. Thus, the firm's value, at an optimum, is

$$V_t = \int_t^\infty \exp\left(-\int_t^s \frac{\rho}{(1-c)(1-\lambda_v p)}\,\mathrm{d}v\right)\left(1 - \frac{\mu_s}{1-\lambda_s p}\right) G_s\,\mathrm{d}s. \tag{2.10}$$

Assuming that $\theta > c$, at least one of the multipliers in Equation (2.9) must be nonzero. At the margin, issuing new shares to pay dividends increases taxes

(the increase in dividend taxes exceeding the reduction in capital-gains taxes) and reduces the value of shares. This cost is reflected by the negative term $\left(\frac{1-\theta}{1-c}-1\right)$ in expression (2.8). To maximize value, firms will wish to decrease both new shares and dividends until at least one of the constraints binds. We many distinguish three regimes, according to whether λ, μ, or both are positive. Though firms might switch among regimes over time, it is helpful to consider first the implications for firms permanently in one regime or another.

2.1. The "traditional" view ($\mu = 0$)

When only the minimum-dividend constraint binds, expression (2.10) reduces to

$$V_t = \int_t^{\infty} \exp\left(-\frac{\rho}{1-(1-p)c-p\theta}(s-t)\right) G_s \, ds. \tag{2.11}$$

According to this expression, the value of the firm equals the present value of its cash flows net of new share issues and dividends, discounted with a before-personal-tax discount rate reflecting an individual tax rate on equity income that is a weighted average of the tax rates on dividends and capital gains. In this regime, a fixed share p of the cash flows from any marginal investment are paid out as dividends and taxed at rate θ, with the remainder being retained and taxed at rate c. This regime has been said to reflect the "traditional" view [see, e.g., Poterba and Summers (1985)], because it includes two "standard" conclusions. The first conclusion is that both dividend and capital-gains taxes raise the corporate discount rate, which equals $\frac{\rho}{1-(1-p)c-p\theta}$. The second is that, at the margin, firms will increase value by investing to the point at which the marginal valuation of a dollar of new investment is one dollar. This last point may be seen by noting that reducing a shareholder's wealth by one dollar and increasing the present discounted value of future cash flows G_s by one dollar leaves the representative shareholder indifferent to the outcome.

2.2. The "new" view ($\lambda = 0$)

When only the share-repurchase constraint binds, expression (2.10) reduces to

$$V_t = \int_t^{\infty} \exp\left(-\frac{\rho}{1-c}(s-t)\right)\left(\frac{1-\theta}{1-c}\right) G_s \, ds, \tag{2.12}$$

a valuation expression that has two striking implications. First, the appropriate discount rate, $\frac{\rho}{(1-c)}$, is unaffected by the tax rate on dividends, regardless of the dividend yield. Second, the net cash flows of the firm are multiplied by the ratio $\left(\frac{1-\theta}{1-c}\right) \leqslant 1$. This

result was called the "new" view of dividend taxation [e.g., Auerbach (1981)], which, of course, it was at the time it first received serious analysis as an alternative to the view laid out above, in analyses by King (1974), Auerbach (1979a) and Bradford (1981).

The intuition underlying the new view is that, with the share-repurchase constraint binding, the firm will neither issue nor repurchase shares. Thus, its marginal source of equity funds will be retained earnings. Likewise, any subsequent cash flows generated by a marginal investment will be paid out fully as dividends – they cannot be used to reduce share issues, which are already zero. Hence, the tax consequences of both current investment and future cash flows differ from the previous case. The tax benefit of avoiding current dividend taxes upon investment reduces both the discount rate and the equilibrium valuation of marginal investment.

Consider, for example, a discrete-time example of a firm's decision whether to invest an additional dollar at date t that yields a gross payoff (after all corporate taxes) of $1+r$ dollars at date $t+1$. The cost of retaining a dollar is reduced by the dividend taxes saved, and increased by the capital-gains taxes on induced share appreciation, q. Because the value of new investment per dollar equals its cost to the shareholder, in equilibrium, $q = 1-\theta+cq$, or $q = \left(\frac{1-\theta}{1-c}\right)$. One period later, this investment plus its return is worth $\frac{q[1+r(1-c)]}{q}$ per initial net dollar forgone, if all earnings are retained. If all earnings in the subsequent period are paid out, then the shareholder receives $1+r$ from the firm, a distribution that forces the shareholder to pay $[1+r]\theta$ in dividend taxes, which will be partially offset by the capital-gains tax avoided through the payment of the dividend, qc. On net, the benefit of the entire transaction is $\frac{[1+r](1-\theta)+cq}{q}$. In either case, as long as tax rates are constant over time, the value per initial dollar invested is $1+r(1-c)$. Thus, the dividend tax rate plays a role in valuing the firm, but does not influence its investment.

Another way to view this equilibrium is that individuals face a tax on capital income of c and a tax on dividend distributions of $\left(\frac{\theta-c}{1-c}\right)$. While the capital-income tax affects the cost of capital, the extra tax on distributions affects only value, constituting, essentially, a capital levy that is analogous to that imposed by the cash-flow component of the consumption tax [Auerbach and Kotlikoff (1987)]. Even though the levy is not assessed immediately, its inevitability causes it to be capitalized into share prices, a result reflected in the theory's alternative characterization as the "trapped-equity" view. Whether the equity really is trapped is, of course, a central question, to which we return below.

2.3. The intermediate case ($\lambda, \mu > 0$)

Between the two regimes just discussed is a regime in which both constraints bind. In this situation, firms pay minimum dividends and issue no new shares. In terms of the value of the firm, this regime is intermediate between the "traditional" and "new" regimes, ranging from the former for small values of μ to the latter for small values of λ. Characterizing the cost of capital is more complicated, because the multipliers may take on a range of values, and the cost of capital depends not only on

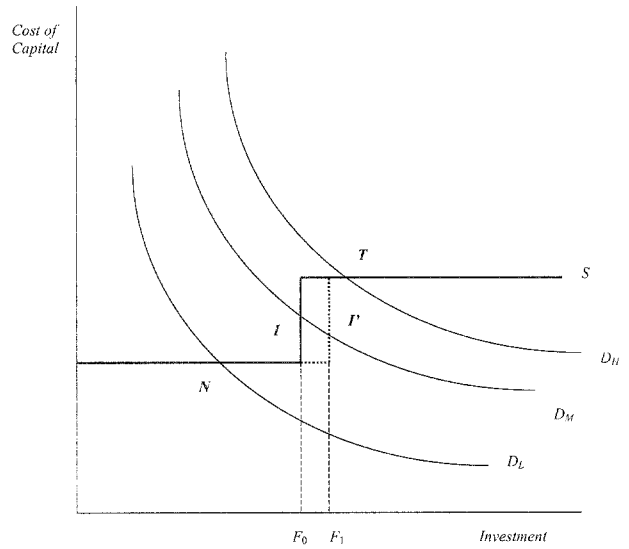


Fig. 1. Alternative equity policy regimes.

the current values of these multipliers, but also on their rates of change[3]. However, for constant values of μ and λ, the cost of capital, also, will lie between those of the two previous regimes.

Which of these regimes is most likely to occur depends upon the relationship between the firm's investment and its cash flow, as is depicted in fig. 1, which illustrates the equilibria corresponding to different investment demand schedules, labeled D_H, D_M, and D_L. The schedule S represents the supply of funds, and reflects the fact that external funds are more costly to the firm than internal funds[4]. For firms with high levels of investment demand relative to cash flow, as represented by demand curve D_H, the maximum level of retained earnings, F_0, is far short of investment, and the equilibrium will be at point $\boldsymbol{T}$, with new share issues needed as a supplementary source. For firms with low levels of investment relative to cash flow, as represented by demand curve D_L, the level of retained earnings available at the dividend constraint exceeds the amount needed to finance investment, so dividends will be reduced until the share repurchase constraint binds, at point $\boldsymbol{N}$. Between these two regimes, as at point $\boldsymbol{I}$, firms will finance all investment through retentions, not

[3] The general expression for the cost of capital is given in expression (3.3) below.

[4] As already discussed, the cost of funds for the intermediate regime will depend not only on the values of μ and λ, but also on these multipliers' rates of change. This will be true, too, for firms transiting out of one of the other two regimes, for which the multipliers will be changing. However, for given future values of the multipliers, the cost of funds will still be higher if external funds are used currently, as in the constant-multiplier cases just analyzed.

finding additional investment profitable enough to justify the more expensive external equity funds, but finding investment profitable enough not to increase dividends and reduce investment. For firms in this regime, a shift in the level of internal funds, say from F_0 to F_1, will increase investment, providing one of the theoretical arguments in support of the observed dependence of investment on cash flow [e.g., Fazzari et al. (1988)].

Although movement among regimes can be driven by shifts in either demand or supply schedules, it is a useful simplification to think of firms with internal funds that are adequate to finance investment as "mature" and those with stronger demand relative to existing funds as "immature". The resulting classification leads to the notion [Sinn (1991a)] of a life-cycle process for firms with respect to equity tax regime, a so-called "nucleus" theory in which the firm begins in the traditional regime and eventually makes the transition to one in which internal funds suffice. This distinction highlights the fact that equity taxation may represent a barrier to entry for new firms [Judd and Petersen (1986)].

2.4. Corporate tax integration

The distinction between these alternative views of the impact of shareholder-level taxes is highlighted by the issue of corporate tax integration, which encompasses a range of policies aimed at alleviating the double taxation of corporate-source equity income. All such policies involve a reduction in taxes on dividends, either at the shareholder level, through a direct reduction in tax rates or through credits or deductions that have the same effect. The two approaches adopted most commonly in practice around the world are the *imputation* system that provides tax credits to shareholders on dividends received and the *split-rate* system that taxes corporate earnings distributed as dividends at a lower rate. Although these systems differ in their details, they are fundamentally equivalent in their incidence and incentive effects, lowering the total tax burden on dividends. Their impact can be analyzed by considering a reduction or elimination of the tax rate on dividends, θ.

The analysis depends on the size of the reduction in the effective tax rate on dividends. If θ is reduced but remains greater than c, the model as presented above applies. Under the traditional view, firms would receive a reduction in their cost of capital; under the new view, only the capital levy would fall, with no impact on the cost of capital. If θ falls below c, then neither of the constraints (2.5) or (2.6) will bind, the associated multipliers will be zero, and the three regimes discussed above collapse to one. This is because firms will now *reduce* taxes and gain from issuing shares to pay dividends. Some additional constraint is necessary to prevent infinite tax arbitrage, and one is typically present in existing systems that restricts the tax relief to dividends attributable to previously taxed corporate earnings. Thus, once all earnings have been distributed, there is no further tax incentive to distribute. In our model, the easiest way to represent this is by a constraint that limits dividends to all earnings,

$$D_t + \dot{V}_t - S_t \geqslant D_t. \tag{2.13}$$

Inserting this constraint into Equation (2.7) in place of the previous two constraints yields, in place of Equation (2.8),

$$V_t = \int_t^{\infty} \exp\left(-\int_t^s \frac{\rho}{(1-c)(1+\gamma_v)}\mathrm{d}v\right) \frac{1}{1+\gamma_s}\left[G_s(1+\gamma_s) + D_s\left(\frac{1-\theta}{1-c} - 1 - \gamma_s\right)\right]\mathrm{d}s, \tag{2.8'}$$

where γ is the multiplier associated with Equation (2.13). Maximizing value with respect to D_s again makes the last term in brackets vanish, with $\gamma_s = \left(\frac{1-\theta}{1-c}\right) - 1$, allowing us to rewrite Equation (2.8′) as

$$V_t = \int_t^{\infty} \exp\left(-\frac{\rho}{1-\theta}(s-t)\right) G_s\,\mathrm{d}s. \tag{2.14}$$

Thus, we can think of integration as having the same impact in all regimes, lowering the cost of capital, once equality between θ and c is reached. Up to that point, integration is less effective at lowering the cost of capital in the traditional regime (because $p < 1$) and not effective at all in the trapped-equity regime[5].

Is there a way to make integration schemes more effective? Because the traditional view applies only to the extent that firms use new shares as their marginal source of equity funds, a tax benefit based on new share issues, rather than dividend payments, would seem to be the answer. One proposal floated in the United States during the 1980s would have given firms a partial deduction for dividends paid (like a split-rate system), with the size of the deduction based on the share of dividends attributable to equity issued after the legislation's effective date[6]. This scheme is basically similar to the "Annell" deduction present in Sweden during the 1970s and 1980s [King and Fullerton (1984), p. 95].

Why aren't such schemes more commonly used? A subsidy to new share issues raises the question of how repurchases should be treated. All firms, even those issuing new shares, would have an incentive to repurchase all outstanding shares and issue new ones to qualify for the deduction. The natural response is to impose a tax on repurchases that offsets any such potential tax benefit. However, taxes on repurchases would hit not only such "churning" transactions, but also transactions by firms engaging in net repurchases. Such activity is, of course, inconsistent with the constraint (2.6), but this constraint reflects a restrictive, simplifying assumption. Repurchases do occur, even if they are less common than tax factors alone would

[5] In a small open economy, another reason why integration may not affect the cost of capital is that the equilibrium rate of return may not be determined by domestic shareholders. For further discussion, see Boadway and Bruce (1992) and Devereux and Freeman (1995).

[6] The proposal was first described in American Law Institute (1982). See Auerbach (1990) for further discussion.

suggest. One important example of repurchasing is the cash-financed takeover in which one firm redeems the shares of its target company. Taxing such transactions would be a controversial policy change.

2.5. Evaluating the models

Researchers have attempted to evaluate the alternative theories of the impact of dividend taxation by testing these theories' implications regarding financial and investment behavior and market valuation. While the predictions appear to differ sharply, testing has proved challenging due both to data limitations and the fact that theories themselves derive from a simplified model that omits certain elements of reality that complicate the interpretation of results.

For example, the first approach one might think of would be to examine the actual patterns of equity finance. We observe, for example, that most firms do not issue new shares in a given year, which would seem to support the new view. On the other hand, if firms face fixed costs of issuing new shares, they might effectively use new issues at the margin by engaging in large, periodic issues. Apparently contradicting the new view is the existence of share repurchases. Repurchases, always present to some extent in the United States, began to grow during the mid-1980s, in concert with the merger wave that occurred at the same time, as firms used cash to purchase the shares of other firms, in addition to their own [Bagwell and Shoven (1989)]. This growth, particularly among large firms, led to the inference that firms finally had "discovered" how to avoid dividend taxation. More recently in the United States, there has been a growth in the percentage of firms not paying dividends [Fama and French (2000)].

However, the implications are not so clear. Note that what is crucial for the new view is the *relative* taxation of the sources and uses of funds. For example, if firms obtain equity funds by reducing repurchases and retaining earnings, and distribute funds by increasing repurchases and dividends in the same proportion, then the new view is essentially intact. All that is needed is to apply a different value of the personal tax rate instead of θ to reflect the fact that some distributions are taxed at rate θ and others are taxed at rate 0 [Sinn (1991b)]. The same logic would apply if firms retained earnings and issued equity to finance investment and used the proceeds of investment to increase dividends and reduce new share issues in the same proportion. Thus, rejection of the new view requires showing not only that dividends are an unimportant marginal source of funds, but also that reducing the issuance of new shares is an unimportant marginal *use* of funds. A piece of evidence on this particular implication is discussed below[7].

[7] Even under the assumption of the traditional view that the firm relies on equity issues as a source of funds but not as a use of funds, the cost of capital may be independent of the dividend tax rate. An example is provided by Bernheim (1991), who develops a signaling model in which the fraction of distributions taking the form of dividends rather than repurchases responds to changes in the dividend tax rate to preserve the average tax rate on distributions.

Moving beyond simple observed patterns of finance, researchers have tested other implications of the alternative theories. In a widely cited paper seen as providing empirical evidence in favor of the traditional view, Poterba and Summers (1985) estimated equations based on Tobin's q-theory of investment. This theory predicts that investment by firms facing convex adjustment costs will be positively related to the relationship between the marginal value of capital, q, proxied by the stock-market value per unit of capital, and the long-run equilibrium value of capital, q^*, i.e., $I = f\left(\frac{q}{q^*}\right)$. Under the traditional view of taxation, with marginal equity funds coming through new share issues, $q^* = 1$. Under the new view, $q^* = \left(\frac{1-\theta}{1-c}\right)$. Using postwar data from the United Kingdom, Poterba and Summers estimated investment equations of the form

$$I = f\left(\omega q + (1-\omega)\frac{q}{(1-\theta)/(1-c)}\right),$$

accepting the hypothesis that $\omega = 1$ but rejecting the hypothesis that $\omega = 0$.

However, this result relies on certain restrictive assumptions. First, the calculation of θ and c requires that one identify the "marginal" investor whose tax rates determine valuation under the new view. Poterba and Summers used average marginal tax rates, a seemingly straightforward approach. Yet the marginal equity investor's identity depends on the nature of financial equilibrium. If, for example, the "Miller" equilibrium to be discussed in Section 3 prevails, then the appropriate values of θ and c are instead those for investors who are just indifferent between debt and equity. Given that identification in the UK sample comes from frequent changes in tax rules affecting dividends, errors in measuring the change in $\left(\frac{1-\theta}{1-c}\right)$ would tend to bias the results in favor of the traditional view. Second, the test is meaningful only if the assumptions of the q-theory itself are satisfied, among them that firms face convex adjustment costs, capital is homogeneous and accurately measured, and returns to scale in production are constant. There has been a continuing dispute about the nature of adjustment costs, and even recent evidence in support of the q-theory using panel data [Cummins, Hassett and Hubbard (1994)] suggests that aggregate measures of q contain considerable noise, and that tests based on these – such as those performed by Poterba and Summers – would be biased.

A second empirical finding often taken to favor the traditional view is that dividend payout ratios respond positively to the return to a before-tax dollar of dividends relative to a before-tax dollar of capital gains, $\left(\frac{1-\theta}{1-c}\right)$. While this evidence certainly supports the argument that taxes influence dividend policy (and therefore contradicts the so-called "tax irrelevance" view based on the hypothetical availability of offsetting tax arbitrage strategies), it is less clearly evidence in favor of the traditional view specifically.

The argument that this evidence is inconsistent with the new view is based on the new view's prediction that the level of dividend taxes has no impact on the incentive to invest or pay dividends. However, there are two distinct reasons why an increase in dividend taxes would reduce distributions under the new view.

First, a *temporary* increase in the dividend tax rate does raise the cost of paying dividends under the new view, for it reduces the opportunity cost of funds more than the ultimate burden on the returns to investment. Indeed, consistent with this logic Poterba and Summers (1985) found (based again on an analysis of UK data) that dividends fall with a current rise in dividend taxes and rise with an anticipated rise in dividend taxes, even when the *level* of dividend tax rates is held constant.

Second, an increase in the dividend tax typically does not occur in isolation. In the United States, for example, dividends and interest are taxed at the same rate for individual investors. An increase in dividend taxes also raises the tax rate on interest income, a change that makes corporate investment more attractive by raising the tax burden on alternative investments. Thus, it should spur more corporate investment and, under the new view, a reduction in dividends.

That the cost of paying dividends may increase with the dividend tax rate even under the new view helps in interpreting related evidence on dividend signaling. In a study that focused on the question of whether tax-based signaling drives dividend policy, Bernheim and Wantz (1995) reasoned that if dividends are used as a signal, their information content should relate to their cost. Hence, the increase in value in response to a unit increase in announced dividends should be higher during periods with a higher tax penalty on dividends. Looking at the period 1978–1988, Bernheim and Wantz estimated that the information content per dollar of dividends fell along with the tax rate on dividends in 1981 and again in 1986. While their measure of the cost of dividends was based on the traditional view, their finding is not necessarily inconsistent with the cost of paying dividends based on the new view: the relevant cost under the new view might well have fallen over time as well. For example, anticipations of reductions in marginal tax rates prior to 1981 and again before 1986 should have raised the opportunity cost of paying dividends relative to the cost after rates had reached historically low values after 1986 and would not have been expected to fall further.

Other evidence, based on micro-data, suggests that neither pure regime applies to all firms, but that some firms appear to behave as predicted by the new view. For the United States, Auerbach (1984) estimated that firms issuing new shares required a higher rate of return on investment than those not issuing new shares, as would be the case if the respective costs of capital of the two groups were those of the traditional and trapped-equity regimes. Bond and Meghir (1994) found a higher sensitivity of investment to internal funds among UK firms with low or no dividends payouts. Auerbach and Hassett (2000) found that new share issues were just as responsive to internal cash flow as to investment among all firms that have paid dividends at some point in their observed history, contrary to a key "traditional view" assumption. With respect to dividend policy, they found that dividends responded more strongly to investment and internal cash flow among US firms with characteristics associated with weaker access to external capital markets.

3. The debt–equity decision

For corporations, interest payments are tax deductible, but returns to equity investors are not. Dividends are subject to double taxation, and even returns to equity in the form of capital gains are subject to at least one level of tax, at the corporate level. Thus, there appears to be a strong tax incentive to use debt to fund the firm's activities.

Consider again the firm's valuation under optimal equity policy, as given in Equation (2.10). Recall that we defined G_t as the net proceeds from the firm's operations before the determination of dividends and new share issues. Let us now divide G_t into those flows before interest and debt, X_t, and those associated with debt, B_t, the latter flows being equal to net borrowing less after-tax interest payments:

$$G_t \equiv X_t + \dot{B}_t - i_t(1-\tau)B_t, \tag{3.1}$$

where i_t is the interest rate at date t. Inserting Equation (3.1) into Equation (2.10) yields:

$$V_t = \int_t^\infty \exp\left(-\int_t^s \frac{\rho}{(1-c)(1-\lambda_v p)}\,\mathrm{d}v\right)\left(1-\frac{\mu_s}{1-\lambda_s p}\right)\left(X_s + \dot{B}_s - i_s(1-\tau)B_s\right)\,\mathrm{d}s. \tag{3.2}$$

Maximizing V_t with respect to B_s yields the first-order condition $\frac{\partial V_t}{\partial B_s} - \frac{\mathrm{d}(\partial V_t/\partial \dot{B}_s)}{\mathrm{d}s} = 0$. Letting $\alpha_s = \left(1-\frac{\mu_s}{1-\lambda_s p}\right)$ be the adjustment term multiplying corporate cash flows at date s, this first-order condition implies that

$$i_s(1-\tau) = \frac{\rho}{(1-c)(1-\lambda_s p)} - \frac{\dot{\alpha}_s}{\alpha_s}. \tag{3.3}$$

The right-hand side of Equation (3.3) is the firm's cost of equity capital at date s, taking account not only of the direct cost of funds but also of the capital gains or losses associated with a shift in equity policy regime[8]. The left-hand side of Equation (3.3) is the net cost of borrowing, so Equation (3.3) calls for the firm to equate the costs of debt and equity. If the equity regime is fixed over time and α does not change, then condition (3.3) simply requires that $i(1-\tau) = \frac{\rho}{1-\phi}$, where $\phi = [1-(1-c)(1-\lambda p)]$ is the effective tax rate on returns to equity, ranging from a value of c when $\lambda = 0$ (i.e., under the new view) to $(1-p)c + p\theta$ when $\lambda = 1 - \left(\frac{1-\theta}{1-c}\right)$ (the traditional view).

For a single, representative household also to be indifferent between debt and equity, it must be the case that the returns after individual taxes are equal, or $i(1-\psi) = \rho$,

[8] This term implies that equity is more costly as the firm makes the transition from the traditional regime, due to capital losses as the valuation of capital assumes its "trapped-equity" level.

where ψ is the individual tax rate on interest income. This yields the following condition for firm optimization:

$$(1-\tau)(1-\phi) = (1-\psi). \tag{3.4}$$

Expression (3.4) has a straightforward interpretation. The left-hand side is the net return to the individual investor of a dollar of corporate source income taxed as an equity return. The right-hand side is what the same dollar would yield if passed through as an interest payment. Note, though, that if all tax rates are given, there is nothing obvious that will cause the equality (3.4) to be satisfied; firms will not achieve an interior solution, and will increase or decrease debt until some other constraint binds. In the apparently likely case that $(1-\tau)(1-\phi) < (1-\psi)$, one would obtain a corner solution with an all-debt outcome.

Some have embraced this argument. Perhaps most prominent is Stiglitz (1973), who suggested that firms should use equity to cover the capitalization of ideas, thereby avoiding immediate capital-gains taxes, but that debt should support any new investment by existing enterprises. However, this prediction seems at variance with the evidence. Though debt–equity ratios have varied across countries and time periods, equity finance has generally accounted for a larger share than debt of corporate capital structures, at least in the aggregate. This section reviews the different theories of corporate leverage, and the associated empirical evidence.

The simplest explanation for why firms don't borrow more is that, at the margin, there exist non-tax costs that offset the tax advantages of doing so. To understand these costs, it is first necessary to clarify the characteristics that distinguish debt from equity, for tax purposes.

According to tax rules in the United States and elsewhere, debt involves a fixed commitment to make payments, while equity does not. Thus, the more debt a firm issues, the greater its commitment of future cash flows to making interest payments, and the greater the probability that its cash flows will be inadequate to cover interest payments. This increases the probability of bankruptcy or other financial distress, the resource costs associated with which would be taken into account when making the initial borrowing decision. That these costs matter to some extent is supported by the efforts made by tax authorities to deny interest deductibility to "debt" for which commitments to pay interest and principal are weakened. In the United States, for example, there are limits on the deductibility of interest on "non-recourse" debt (for which creditors literally have no recourse if payments are not made) and on very long-term debt, for which principal repayment is of little concern.

A second possible non-tax cost to borrowing derives from the information asymmetry between potential lenders and borrowers. In an environment where lenders cannot distinguish between good and bad risks, adverse selection may occur, as firms that are relatively less risky will be discouraged by the large risk premium imposed by lenders, and only the bad risks will find borrowing attractive [Stiglitz and Weiss (1981)].

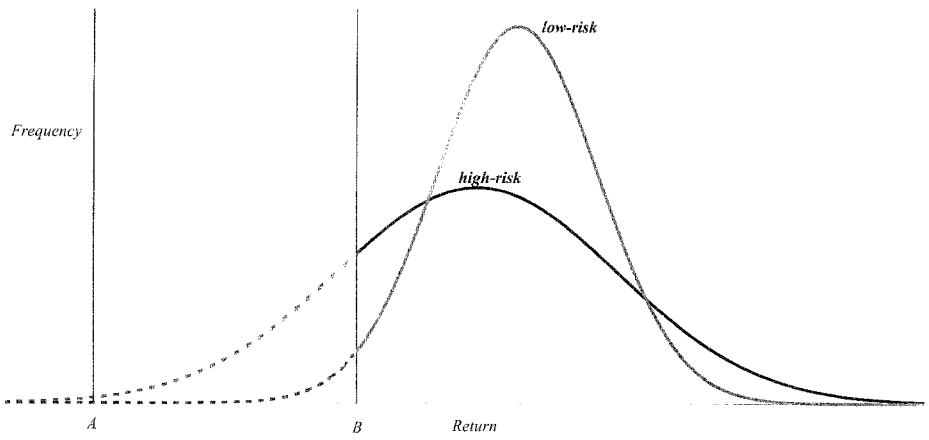


Fig. 2. Borrowing and moral hazard.

Yet another imperfect-information explanation relates to the moral hazard problem of firms that can alter their investment choices to take advantage of debt-holders. With limited liability, firms face a one-sided bet when making risky investments: if their investments fail, the downside risk is truncated. As illustrated in Figure 2, which depicts two possible return distributions, any return below that labeled A would induce bankruptcy. Increased leverage raises the position in the return distribution at which failure occurs, as shown in the figure in the move from point A to point B. But the impact differs according to the underlying risk of the firm's assets. The riskier the firm's assets, the greater the share of the distribution that will be truncated by the shift. Hence, firms may be encouraged to undertake riskier investments, to take greater advantage of limited liability. In Figure 2, this might make the high-risk investment more attractive than the low-risk investment, despite its full distribution having a lower mean return. One may view the ability to walk away from losses as a put option that creditors provide as part of the lending contract. Undertaking riskier investments increases the value of this put option. Creditors, of course, would charge for this put option were the firm's investment strategy known and fixed, but such a "wealth transfer" cannot generally be avoided otherwise. The more difficult it is to monitor a firm's activities, and the easier it is for a firm to alter its asset portfolio, the more of a problem this potential moral hazard imposes and the higher a premium lenders would insist on [Myers (1977)]. The associated inefficiency in the choice of investment projects would thus be impounded in the cost of borrowing.

While each of the previous explanations relates to why value-maximizing firms might limit their borrowing, managers might well stop short of optimal borrowing because of a divergence between their incentives and those of their shareholders. Managers with high debt loads might well be forced to work harder, their human capital at considerable risk should the firm be forced into bankruptcy. Though this effect increases the attractiveness of debt from the shareholder's perspective, it has

quite the opposite impact on managers, who would find the prospect of considerable "free cash flow" much more enticing [Jensen (1986)].

In addition to these theories of borrowing, there are others that relate more directly to the apparent tax incentives themselves. In responding to the apparent inequality in favor of debt finance – $(1-\tau)(1-\phi) < (1-\psi)$ – the theories suggest that (1) ϕ is not as large as one might think; (2) τ may be smaller than the statutory corporate tax rate; and (3) ψ may be much larger than ϕ for the relevant individual investor. The first of these arguments follows from the new view of equity taxation discussed above. From that perspective, shareholders face only the capital-gains rate on marginal equity returns, even those that flow in the form of dividends. If $\phi = c$, and c is very small, then the debt–equity decision rests roughly on the relative magnitudes of τ and ψ, which may not be far apart. The other two explanations, which we now explore in more detail, are that the ability to deduct interest payments may be limited, and that the relevant marginal investor is one for whom the corporate tax advantage for debt is offset by a personal tax advantage for equity.

3.1. Competing tax shields

The absence of a unique interior optimum in simple models of the debt–equity decision follows from the fact that tax rates are assumed not to change with the debt–equity ratio. Thus, if the inequality $(1-\tau)(1-\phi) < (1-\psi)$ holds at a low debt–equity ratio, it will hold at higher debt–equity ratios and continue to encourage borrowing. This result requires that interest payments be deductible at the corporate tax rate, τ, regardless of their magnitude. But corporate tax rules do not conform to this assumption. Instead, they limit deductions for interest and other expenses to the extent that these deductions would induce negative taxable income and tax refunds. That is, if the corporation's earnings before interest and taxes, or EBIT, are E, and its interest deductions are I, then the tax system treats positive and negative values of $(E-I)$ asymmetrically,

$$T = \begin{cases} \tau(E-I) & \text{if} \quad (E-I) > 0, \\ \tau^*(E-I) & \text{if} \quad (E-I) < 0, \end{cases} \tag{3.5}$$

with $\tau^* < \tau$[9]. The simplest such asymmetry is that $\tau^* = 0$ – no deductibility for losses – but tax systems typically provide *some* tax benefit even for firms with losses through the ability to carry losses forward or backward to other tax years. We discuss below how one estimates the value of such unused current deductions.

The likelihood that a firm's interest payments exceed its EBIT depends not only the debt–equity ratio, but on other elements of the tax system as well. If the tax system

[9] For a multinational corporation, additional limits may apply. In the United States, for example, only a portion of the interest on domestic borrowing may be used to offset domestic source income. See Froot and Hines (1995).

measured a corporation's income accurately then, in a riskless world, it would be possible to finance all investment by borrowing and just deduct all interest payments. To see this, consider the derivative of the valuation expression (3.2) with respect to time, t:

$$\frac{\rho}{(1-c)(1-\lambda_t p)} V_t = \left(1 - \frac{\mu_t}{1-\lambda_t p}\right)\left(X_t + \dot{B}_t - i_t(1-\tau)B_t\right) + \dot{V}_t. \tag{3.6}$$

If the firm finances all of its operations by borrowing, then it keeps its equity value exactly at zero, i.e., $V_t = \dot{V}_t = 0$. In this case, the equilibrium valuation condition (3.6) becomes

$$i_t(1-\tau)B_t = X_t + \dot{B}_t, \tag{3.7}$$

which says that the return to debt equals the firm's real net cash flows, X_t, plus the additional amount of debt the firm is able to issue without reducing its equity value, i.e., the increase in the value of the firm. But this is simply the firm's economic income, say E', less taxes computed before interest deductions, τE. Thus, we may rewrite Equation (3.7) as

$$i_t(1-\tau)B_t = E' - \tau E, \tag{3.8}$$

from which it follows that interest payments will be less than, greater than, or equal to EBIT, E, according to whether E is less than, greater than, or equal to economic income, E'. Indeed, if economic income and EBIT are equal, then we may cancel the corporate tax rate τ from both sides of Equation (3.8), meaning that the path of the firm's debt, and hence its value, is independent of the corporate income tax [Samuelson (1964)].

In general, though, the corporate tax base deviates from true economic income, as corporate tax systems treat certain types of income – such as corporate capital gains – favorably, thereby lowering the value of E. The same effect is provided by schemes that provide generous deductions for other corporate expenses, notably depreciation. Hence, corporations may well hit the limit of current deductibility at considerably less than an all-debt capital structure, even before account is taken of the fact that *ex post* returns are risky and may fall short of their certainty-equivalent value. Taking risk into account, the tax system's asymmetry described in Equation (3.5) will impose a greater disincentive to borrow on firms with more uncertain returns.

The resulting financial equilibrium, then, will be one in which the equality (3.4) is established by the endogeneity of the corporate tax rate. The statutory tax rate τ is replaced in the equation by a function of τ and τ^* that takes into account both the likelihood that the firm will not be able to deduct marginal interest payments immediately and the value of such deferred deductions. This equilibrium, of course, will also be affected by the risk and tax characteristics of the assets in which the firm

invests. The situation presents the firm with a trade-off between interest deductions and other tax deductions, as explored initially by DeAngelo and Masulis (1980) and analyzed in more detail by Sinn (1987). While, *ceteris paribus,* firms would generally seek to maximize other deductions, they may not do so if there are direct costs involved (as through a distortion of investment decisions) or if there are other advantages to borrowing, such as monitoring that debt-holders may provide [Kanniainen and Södersten (1994)][10].

3.2. The Miller equilibrium

For many tax systems, the corporate tax rate is well above the average marginal tax rate on interest income. In the United States, the corporate tax rate at the turn of the century was 35 percent, while the highest marginal tax rate (subject to small further adjustments) was 39.6 percent. With a substantial share of assets held by tax-exempt institutions such as pension funds, for whom only the corporate tax on equity income applies, it seems clear that the typical investor would face a lower total tax burden on debt than on equity.

But, as elaborated by Miller (1977), how one defines the relevant marginal investor depends on the nature of financial equilibrium. In a world in which investors choose to hold only debt or only equity according to which yields a higher after-tax return, all that is necessary for an interior solution is that there exist *some* investors who prefer equity (and some who prefer debt) for tax purposes. This equilibrium is illustrated in Figure 3, which plots the relative personal tax preference for debt, defined by the ratio $\frac{1-\psi(y)}{1-\phi(y)}$, as a function of income, y, along with the corporate tax preference for equity, $1-\tau$ (which is independent of an individual's income). If marginal tax rates increase with income, and the individual tax on equity, ϕ, is some fraction of the tax rate on debt, ψ, the tax preference for debt will be decreasing in y, as shown in the figure. At income level y^*, the two curves cross and expression (3.4) is satisfied.

Clearly, if all investors had income y^*, debt and equity would be equally preferred and firms indifferent in equilibrium. But even with a range of investors with incomes below and above y^*, there will still exist an equilibrium in which equity and debt coexist and firms are indifferent between them. Firm indifference alone does not require that expression (3.4) hold, merely that the required return to equity, ρ, equal the after-tax interest rate $i(1-\tau)$. Assuming this condition to be met, we can see that those for whom $y > y^*$ will receive a higher return from holding equity than from holding

[10] Kanniainen and Södersten derive their result in the context of a model in which firms face the "one-book" accounting constraint used by some countries, in which dividends can be paid only out of taxable earnings. In this model, an increase in non-interest tax deductions requires a reduction in dividends and hence in borrowing. Further discussion of the implications of one-book accounting on financial decisions and the cost of capital may be found in Kanniainen and Södersten (1995). Also see Sørensen (1994), who provides an integrated discussion of this constraint and those underlying the equity policy regimes discussed above.

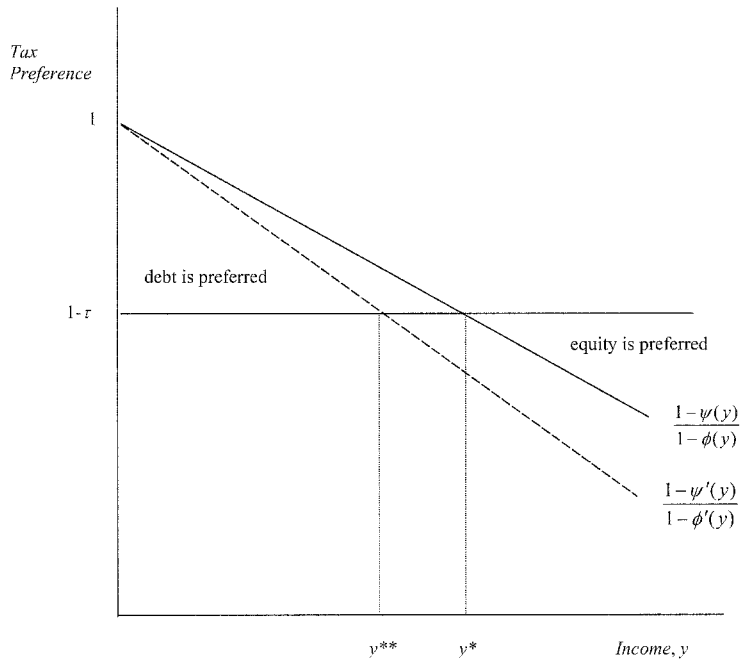


Fig. 3. The Miller equilibrium.

debt, and those for whom $y < y^*$ will receive a higher return from holding debt than from holding equity. Thus, if individuals may hold only positive quantities of either asset, then the market for debt and equity will clear if firms – who are indifferent with respect to their individual debt–equity choices – issue just enough debt to satisfy the demands of those with incomes below y^*. Hence, only the aggregate debt–equity ratio, and not that of any firm, will have a determinate solution.

This theory characterizes the marginal investor not as some "representative" investor, but rather as the investor who is indifferent from a tax perspective between debt and equity. Thus, if the tax system changes, the identity of the representative investor will change, too. For example, if the tax system shifts in favor of equity (as would be the case if individual tax rates rose) then, as illustrated by the dashed schedule in Figure 3, the marginal investor would become someone with lower income, y^{**}, and the aggregate corporate debt–equity ratio would fall.

The Miller model is easily generalized to the case of more than two types of assets, for example, with the addition of completely tax-exempt municipal bonds in which the very-highest-bracket individuals would specialize. But the model confronts a more serious limitation, namely that its prediction of investor specialization is patently false. In the real world, even non-taxable pension funds hold a substantial share of their assets in the form of equity. This contradiction arises because the Miller model presumes that assets differ with respect to tax treatment alone, so that there is no

trade-off with respect to other characteristics. In a more general model in which debt and equity differ with respect to risk as well, Auerbach and King (1983) showed that the Miller result generally requires the asset space to be sufficiently complete to permit "tax spanning" that lets households choose return patterns and tax treatment separately. Otherwise, households will hold portfolios diversified with regard not only to individual equity holdings, but also with regard to debt and equity[11]. With such diversification comes a redefinition of the "marginal" investor. Now, the tax rates of all investors holding both debt and equity matter in the calculation, entering in a weighted average. The weights depend on the degree of absolute risk aversion, the less risk-averse individuals[12] in a better position to arbitrage differences in rates of return playing a more powerful role.

3.3. Evidence on the effects of taxation on corporate borrowing

One seemingly obvious approach to evaluating whether tax rules favor borrowing is to estimate the impact of changes in debt on firm value. However, a little thought reveals why such an approach is unlikely to succeed. Because the change in a firm's debt does not result from a random process, the market's response reflects not only its valuation of the change itself, but also whatever information the change conveys. As noted by Fama and French (1998), these effects are difficult to separate. Moreover, even if adequate controls for information effects did exist, valuation responses would merely reveal the presence of deviations from optimal policies, rather than the underlying influence of taxes. That is, for any model in which firms eventually settle at an interior optimum, either because marginal tax benefits decline or marginal non-tax costs increase, the marginal impact on value of a change in debt should be zero. Positive responses to increases in debt would suggest that firms had initially settled on debt–equity ratios that were too low, and negative responses would suggest that initial debt–equity ratios were too high, with neither outcome revealing anything about the size of the tax benefits at any given level of borrowing. While the pure Miller model would predict no valuation responses (controlling for information effects) because optimal firm policy is indeterminate, the lack of a measured response might also simply reflect that firms, on average, are at their respective unique optima. Such an exercise, then, might shed light on whether managers act in the interests of shareholders, rather than telling us much about the tax benefits of leverage.

Most empirical investigations of the importance of tax rules with respect to the choice of financial structure may be classified into two main categories. The first group of studies estimates the extent to which interest payments are tax deductible, shedding

[11] In his chapter (17) in this Handbook, Poterba considers the portfolio-choice implications of the Auerbach–King model in more detail.

[12] With well-behaved preferences toward risk characterized by declining absolute risk aversion, individual weights would be increasing not only with respect to risk tolerance, given wealth, but also with respect to wealth.

light on the potential importance of competing tax shields as an explanation of limited borrowing. The second empirical approach has been to estimate models of leverage decisions using cross-section or panel data, including tax and non-tax characteristics of firms to assess the relative importance of tax factors. Except where noted below, this literature takes little account of the personal tax considerations relevant to Miller's explanation of financial policy.

3.3.1. Limits on interest deductions

As discussed above, tax systems typically provide less than full loss offset, not giving a tax refund to those investors with negative current taxable income. However, this does not imply that prospective incremental interest deductions have no value in such circumstances. First of all, firms that borrow do not necessarily know, *ex ante,* that they will have negative taxable income in a given year. One would wish to weight the value of interest deductions in any state by the probability of that state occurring, evaluated at the time of the borrowing decision. Second, even if interest deductions cannot be taken immediately, this does not mean that they can never be used. Rather, unused deductions typically can be carried forward for possible use in a subsequent year and, in some countries, carried back to a prior tax year. For several years in the United States, including the period considered by the research discussed below, the carry-forward period was 15 years and the carry-back period 3 years[13].

Carrying deductions forward reduces their value, because deductions carried forward do not earn interest and may expire unused. Carrying deductions back (by recomputing a prior year's tax liability) produces an immediate deduction. However, the existence of a carry-back provision complicates calculations because it attaches an option value to taxable income, associated with the possibility that the firm may wish to carry future losses back to the current year. This, in turn, reduces the value of an immediate deduction when the firm *is* taxable.

To solve for the value of interest deductions in this environment, some assumptions are necessary. Imposing the restriction that firm transitions between taxable and non-taxable states follow a second-order Markov process, Auerbach and Poterba (1987) derived an algorithm to solve for the present-value tax liability associated with a dollar of taxable income. (This calculation also measures the value of a one-dollar reduction in taxable income due to an interest deduction). Altshuler and Auerbach (1990) extended this methodology to take account of intermediate states in which firms may deduct some but not all expenses[14]. The general methodology of these two papers can be understood by considering a simplified case in which transitions follow

[13] Currently in effect are the provisions of the Taxpayer Relief Act of 1997 that reduced the carry-back period to 2 years and increased the carry-forward period to 20 years.

[14] A related intermediate state arises in the case of the alternative minimum tax (AMT), under which a firm faces a marginal tax rate below the statutory corporate rate. See Lyon (1990) for further discussion of the effects on incentives of transitions involving the AMT.

a stationary first-order Markov process between two states (taxable and non-taxable) and losses may be carried back only one year. In this case, the "shadow" value (in terms of reduced taxes) of a dollar to be deducted (or the cost of a dollar of extra taxable income) is the statutory tax rate multiplied by

$$\begin{aligned} w &= \sum_{i=1}^{L} \beta^i \pi_{\mathrm{NN}}^{i-1} \pi_{\mathrm{NT}}(1-\upsilon) \quad \text{in state N,} \\ 1-\upsilon &= 1-\beta\pi_{\mathrm{TN}}(1-w) \quad \text{in state T,} \end{aligned} \tag{3.9}$$

where N is the non-taxable state, T is the taxable state, β is the one-year discount factor, L is the number of years after which loss carry-forwards expire, and π_{ij} is the transition probability from state i to state j.

The first of expressions (3.9) says that the value of a dollar tax deduction for a firm not currently taxable is based on the distribution of dates when that deduction can first be used. The probability of its use one year hence is π_{NT}; the probability of its use two years hence is $\pi_{\mathrm{NN}}\pi_{\mathrm{NT}}$; and so on. Payments at each future date must be discounted and adjusted by the term υ to account for the fact that reducing taxable income also reduces the option value of subsequent carry-backs. The second of expressions (3.9) says that a dollar deduction when taxable has its value reduced by the extent to which it precludes subsequent carry-back, the value of which, in turn, is the difference between immediate use and eventual use, $(1-w)$.

Using US corporate tax returns from the period 1970–82 to estimate transition probabilities, Altshuler and Auerbach estimated 1982 shadow values of marginal interest deductions ranging from 19 percent for firms with two successive years of tax losses to 39 percent for firms with two successive years facing no tax constraints. Their asset-weighted sample average was 32 percent, well below the statutory corporate rate of 46 percent prevailing at the time. Thus, the calculations suggested that tax asymmetries were quantitatively important for the corporate sector as a whole and that there was also considerable heterogeneity with respect to the value of interest deductions.

More recently, an alternative approach has been to simulate distributions of tax payments using a large number of random draws based on the assumption that a firm's taxable income follows a random walk. Doing so, Graham (1996) estimated a slightly lower mean value (30 percent) for 1982 than Altshuler and Auerbach for an unweighted sample of COMPUSTAT firms, but a higher value (40 percent) weighting by market value. The gap between the weighted estimates of these two studies may be attributable not only to methodological differences, but also to weighting scheme (market value weights placing more weight on successful firms than asset weights) and also perhaps to sample differences. Altshuler and Auerbach found that their estimates of the incidence of tax losses was higher in actual tax returns than in the corresponding COMPUSTAT records considered by Auerbach and Poterba. For the last year in his sample, 1992, Graham's unweighted and value-weighted estimates of

the average marginal tax rate were 20 percent and 28 percent, respectively, compared to that year's statutory rate of 34 percent[15].

3.3.2. Behavioral responses to variations in tax incentives to borrow

Evidence on the deductibility of interest payments suggests that limits on deductibility have a potential role in explaining observed borrowing decisions. But whether these limitations, or other tax considerations, actually do matter is another question, to be resolved through empirical analysis of the relationship between borrowing and tax incentives.

Implementing a model of borrowing decisions confronts several problems, with which the literature has dealt to varying degrees. First, as just discussed, the tax rate the firm faces on its marginal interest deductions is a complicated function of the firm's current and expected future circumstances. Second, the tax rate at which interest can be deducted is endogenous; the greater a firm's debt, the lower its effective marginal tax rate on interest deductions. Thus, the relationship between borrowing and marginal tax rates based on a simple regression will be biased downward. Third, borrowing may also result from factors correlated with tax status. For example, a firm in financial distress may borrow more as a result, and may also have unused tax credits and deductions. This, too, would impart a downward bias to the relationship between borrowing and the corporate tax rate. Empirical studies typically include other explanatory variables to control for this, some more fully than others. Fourth, there are many different kinds of debt, and close substitutes for debt, such as leases. If only some elements of this category are considered, then the impact of taxation on borrowing as a whole may be misstated. Fifth, measurement of relevant aggregate personal tax rates is difficult, as discussed above in the context of testing theories of the effects of dividend taxation, and measuring shareholder tax-rate variation across firms is even more problematic.

Early empirical work dealt implicitly with the problem of tax-rate endogeneity by using variables that did not depend directly on current debt levels. For example, Bradley et al. (1984) used a proxy for non-debt tax shields equal to the sum of annual depreciation charges and investment tax credits divided by the sum of annual earnings before depreciation, interest and taxes[16]. In cross-section regressions of averages for the period 1962–1981, they found that debt was a *positive* function of the non-debt tax shields, contrary to the theory. In a subsequent cross-section study based on debt averaged over the period 1977–1979, Titman and Wessels (1988) used a factor-analytic

[15] Graham's algorithm also takes into account the AMT that applied during the later years in his sample.

[16] This measure is unorthodox, as it adds together deductions and credits with no tax-rate adjustment, which the authors defend on the basis of not knowing the tax rate to use for such an adjustment. It is hard to know the extent to which the paper's counterintuitive results with respect to this variable are due to its novel construction.

approach to allow their model to define non-debt tax shields as a linear function of three observable measures, including depreciation deductions and investment tax credits (ITCs). They found that tax shields so defined do have the correct sign in predicting long-term debt, short-term debt, and convertible debt in separate equations. However, none of the estimated effects were statistically significant.

Though excluding interest payments themselves, these estimates of non-debt tax shields may be endogenous, as they depend on firm investment choices made simultaneously with borrowing decisions. In cross-section analysis, there is little one can do about this endogeneity, but panel data offer more options. Auerbach (1985), using a panel of firms from 1969 to 1977, attacks the problem in two ways. First, the paper includes fixed firm effects as explanatory variables, to eliminate cross-firm variation in the tendency to borrow that may be correlated with other explanatory variables. Second, it models the *change* in debt–assets ratios rather than their level, and uses a lagged measure of tax capacity – the firm's tax-loss carry-forward – as a measure of the tax incentive to borrow. Estimates of this variable's impact are negative and statistically significant for all borrowing aggregated together, and for long-term borrowing considered separately (but insignificant for short-term borrowing)[17].

MacKie-Mason (1990) adopts a related approach, looking not at changes in debt, but at new public issues of debt relative to new equity issues. While this approach does not control for unobservable firm effects, it does take into account the simultaneous determination of contemporaneous tax and borrowing variables. MacKie-Mason measures tax status by variables used in the previous studies, the tax-loss carry-forward and the investment tax credit. However, he notes that the extent to which the latter variable matters should depend on how close the firm is to tax exhaustion. Thus, he interacts the ITC with a variable meant to measure financial condition, the argument being that the ITC should matter more for firms in poorer condition. As theory would predict, he finds that both terms reduce the probability of issuing debt, with the effects both statistically significant and economically important.

Graham (1996) carries MacKie-Mason's insight about the varying importance of non-debt tax shields one step further, using the methodology discussed in the previous section to estimate each firm's marginal tax rate based on projections of taxable income using each year's initial conditions. He then considers changes in debt as a function of this and other variables and finds that the marginal tax rate exerts a significant effect in pooled data, but is not always significant in individual cross sections. The effect is weakest in 1986 and 1987, around the time of the comprehensive Tax Reform Act of 1986, suggesting the confounding effects of additional factors during this period.

All of the papers discussed thus far limited their attention to firm-level tax incentives, ignoring variations in individual taxes over time and across firms. More recently, Graham (1999) extended the analysis of his earlier paper to include inter-firm variation

[17] In a related context, tax loss carry-forwards are significant in explaining variations in the share of tax-exempt debt in bank portfolios [Scholes et al. (1990)].

in personal tax rates, as well as time-series variation associated with changes in the tax law. In a decomposition of his regression results, he found that only the cross-section ("within") variation, and not the time-series ("between") variation in tax rates exerted a significant impact on his results. Measuring the net tax advantage to debt, in terms of the notation used above [Equation (3.4)], as

$$(1-\psi)-(1-\tau)(1-\phi), \tag{3.10}$$

where ψ is the investor tax rate on debt, τ is the corporate tax rate, and ϕ is the investor tax rate on equity, he assumed identical underlying investor tax rates and achieved cross-section identification through variation in dividend payout rates, basing ϕ on its "traditional" view measure as a weighted average of dividend and capital-gains tax rates. The results support the inclusion of this variable and the implication that personal tax rates do matter, and generally hold up when account is taken of investor clienteles using Auerbach's (1983b) estimates based on ex-dividend-day share price behavior.

Thus, at least in cross-section analysis, firms do seem to respond to differences in tax incentives to borrow. But aggregate responses in time series are less evident. One problem, as identified above in discussing the Miller equilibrium, is that it is not clear how one should aggregate tax rates of different investors. When personal tax rates change, changes in the net tax advantage of debt over equity given in Equation (3.10) will differ across investors in size and even, perhaps, in sign. One method of teasing out the effects of time-series tax-rate variation is to focus on the relative impacts of particularly large tax-law changes on different types of firms, adopting the so-called "natural experiment" approach. An ideal such tax change was the Tax Reform Act of 1986 (TRA86), which lowered tax rates on dividends, interest, and corporate income, raised tax rates on capital gains, and eliminated non-debt corporate shields by repealing the investment tax credit and reducing the acceleration of depreciation allowances on real estate. While the aggregate response of debt to these changes appeared rather small [Gordon and MacKie-Mason (1990)], the changes in debt across firms between 1986 and 1987 do vary as predicted with respect to dividend yield and changes in corporate tax shields [Givoly et al. (1992)].

Another potential source of variation that an investigator might utilize, across countries, has thus far proved difficult to use in isolating the effects of tax factors from other differences. For some initial thoughts in this area, see Rajan and Zingales (1995). In the international context, though, there are other financial margins on which a multinational firm operates. Beyond the choice between domestic issuance of debt and equity, such firms also may decide whether and how to finance abroad. There is considerable evidence that these choices do respond to tax incentives, as discussed by Gordon and Hines in their chapter in this Handbook (Volume 4, forthcoming).

3.4. Leasing as a form of borrowing

As discussed in the introduction, debt and equity are useful simplifications, but financial decisions relate to the allocation of claims to underlying income streams,

rather than to how these claims are packaged and what the packages are called. There may be alternative ways effectively to increase debt without an explicit increase in borrowing. The firm's choice between leasing and purchasing capital provides an illustration of this distinction.

Imagine a firm that plans initially to purchase a unit of capital and finance it entirely with debt, perhaps using the capital as security for the debt. An alternative would be for the firm to lease the capital from another firm, its lease payments to the lessor substituting for its payments of interest and principal. As there appears to be little real distinction between these two situations, this suggests that one might wish to include leases along with explicit debt in assessing the firm's overall leverage. Indeed, as the tax treatment of leases provides one of the key distinctions between borrowing and leasing, we should expect leasing to substitute for borrowing in response to tax considerations.

Consider the decision of whether to lease or purchase a unit of capital that, for simplicity, depreciates exponentially over time. According to the standard Hall–Jorgenson expression for user cost of capital [see Auerbach (1983a)], the zero-profits condition for the owner of this capital is that it deliver a gross (before depreciation and taxes) marginal product equal to

$$C = \frac{(r - \pi + \delta)(1 - k - \tau z)}{(1 - \tau)} \tag{3.11}$$

per dollar of capital, where r is the firm's cost of funds, π is the inflation rate, δ is the rate of economic depreciation, k is the investment tax credit, z is the present value of depreciation allowances per dollar of initial purchase, and, as before, τ is the corporate tax rate. This derivation assumes that the capital owner is always taxable at rate τ and makes use of all available deductions.

In a competitive spot market for capital rental, Equation (3.11) also defines the equilibrium lease payment that such capital should command. For the lessee, then, the lease itself should have no tax consequences, as the lease payment is deductible from the return to capital, leaving a net tax liability of zero. The issue of whether to lease, therefore, hinges simply on the tax consequences of direct ownership. By the assumptions used in constructing the user cost in Equation (3.11), a fully taxable firm facing the tax rate τ will be indifferent between leasing and owning. Thus, a preference for leasing or owning must result from one of these two conditions being violated. In this event, the decision will also depend on the other parameters in Equation (3.11). To see how, we may consider an illustrative example, drawing on the insights of several papers on the subject, including Myers et al. (1976), Brealey and Young (1980) and Edwards and Mayer (1991).

As a benchmark, consider the case in which the inflation rate is zero, there are no investment tax credits, and depreciation allowances are based on economic depreciation. Then Equation (3.11) simplifies to

$$C = \frac{r}{(1 - \tau)} + \delta. \tag{3.12}$$

In this case, depreciation deductions equal δ (the actual rate of decay per dollar of capital), so the tax base after deducting depreciation is simply $\frac{r}{(1-\tau)}$. If the investment is entirely debt-financed, then the cost of funds r equals the after-tax interest rate, $i(1-\tau)$, and interest deductions equal i per dollar of capital. Thus, net taxable income is $\frac{i(1-\tau)}{(1-\tau)} - i = 0$. This outcome corresponds to the "Samuelson" case discussed above, for which deductions for depreciation and interest exactly equal the user cost and there is no tax difference between leasing and owning, regardless of the firm's tax status or tax rate. It is illustrated in panel A of Figure 4, which shows the stream of depreciation deductions, interest deductions, and marginal products of capital – user costs – over time, as the asset depreciates. The sum of depreciation deductions and interest deductions equals the user cost at each date.

Note, however, that if the firm chose not to finance capital entirely with debt, this would reduce interest deductions, leaving the firm with net taxable income at each instant, as shown by the gap between the user cost and the dotted line in the figure[18]. Thus, with economic depreciation and no inflation, leasing rather than buying would not serve to reduce a firm's tax deductions, as it might wish if such deductions could not be utilized. Indeed, with anything short of full debt finance, leasing would appeal only to companies with adequate taxable income.

In reality, depreciation allowances generally are accelerated relative to economic depreciation, and this changes the result just derived. As shown in panel B of Figure 4, accelerated depreciation has two effects. First, it lowers the user cost of capital by increasing the present value of depreciation allowances, z, in expression (3.11). Second, it increases depreciation deductions early in the period of capital use. Together, these effects make it more likely that the sum of interest deductions and depreciation allowances will exceed the user cost, particularly in early years of the asset's life. As unused tax deductions from these years carry forward without interest, this change makes ownership generally less attractive for a company facing tax "exhaustion". The result is even stronger for assets also receiving an initial investment tax credit on top of the large initial depreciation deductions, which also are typically the short-lived assets for which depreciation deductions themselves – and hence the benefits of acceleration – are more important. Thus, with realistic depreciation provisions, firms facing tax limitations are more likely to lease, particularly assets that are short-lived.

Finally, consider the impact of inflation, which has two additional effects. First, because depreciation allowances are not indexed for price-level changes, they fall in real terms over an asset's lifetime. Panel C of Figure 4 illustrates this effect. This reduction in the present value of depreciation deductions, z, also increases the user cost and hence the equilibrium marginal product of capital, so both the direct

[18] This shift in the source of funds may also have an impact of the cost of funds and hence the user cost, depending on the specification of financial policy equilibrium. However, this additional complication does not affect the main conclusion that reduced borrowing increases taxable income.

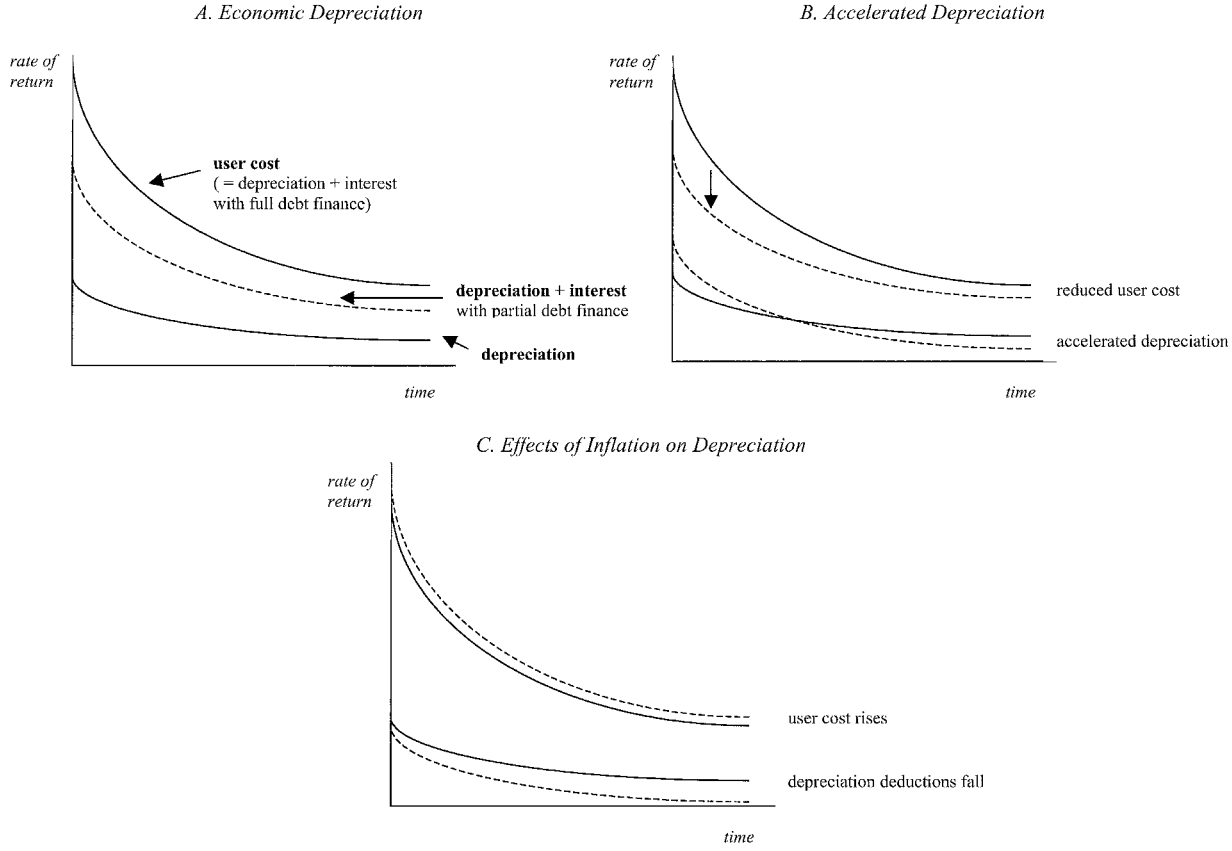


Fig. 4. The leasing decision.

impact on deductions and the indirect impact on the user cost increase taxable income, making direct ownership more attractive to the tax-constrained firm. On the other hand, because nominal interest payments are tax deductible, inflation-induced higher nominal interest rates increase the real value of interest deductions and reduce the required marginal product of capital, pushing the lease-versus-own decision in precisely the opposite direction. The net impact is ambiguous, but clearly depends on the relative importance of depreciation and interest. Inflation is most likely to encourage leasing by tax-constrained firms of assets that are financed largely by debt and assets that are long-lived. The second of these effects counteracts the impact of accelerated depreciation that makes short-lived assets the better candidates for leasing.

In summary, then, the only clear result is that leases will be most attractive for tax-constrained firms – rather than for fully taxable firms – for assets capable of being financed by high proportions of debt. But, as the asset characteristics that facilitate borrowing are likely to be similar to those that make leasing feasible (e.g., homogeneity, adequate resale market, etc.), the common notion that leasing should be done by tax-constrained firms may be reasonable, after all.

This analysis has two implications for empirical work. First, leases may well represent a form of disguised debt that belongs in the construction of estimated debt–asset ratios. Second, leases have different tax characteristics than traditional debt, and may be an attractive substitute for firms facing low effective tax rates. Thus, leases should respond differently than explicit debt to changes in the firm's marginal tax rate. This result is confirmed by empirical evidence relating leasing to the tax-loss carry-forward [Barclay and Smith (1995), Sharpe and Nguyen (1995)], to a more sophisticated measure of the firm's marginal tax rate [Graham et al. (1998)], and to the limitations on interest deductions by multinational corporations induced by interest allocation rules [Froot and Hines (1995)].

4. Organizational form and ownership structure

To this point, we have considered the financial decisions of a given corporation, taking its organizational form and ownership structure as given. However, each of these aspects of the firm may change significantly, even over a very short period of time. This section considers the related topics of changes in organizational form – between regular corporate status and available alternatives – and changes in ownership structure – mergers, acquisitions and spinoffs – in each case considering how these changes may be induced by tax incentives.

4.1. The choice of organizational form

Firms need not exist as corporations. In the United States, for example, master-limited partnerships and "S" corporations (named for a location in the tax code and as distinguished from the traditional "C" corporations) permit multiple equity holders

and preserve the traditional corporate benefit of limited liability. Yet the earnings generated by each form of enterprise are "passed through" and subject to taxation only of individual owners, with no additional tax imposed at the entity level. Thus, the tax incentives affecting the choice between corporate and non-corporate status resemble those between debt and equity.

As with the debt–equity decision, recent configurations of corporate and individual tax rates appear to point in favor of taxation at the individual level only. However, there are a number of other aspects of the corporate–non-corporate choice that distinguish it from the debt–equity choice already analyzed. First, the choice is a discrete one for any given firm, which cannot choose to be "partially" corporate[19]. However, heterogeneity among firms will still generally lead to interior solutions with regard to corporate status for any given industry or group of firms.

Second, the non-tax costs of opting out of the corporate form differ from those of borrowing. While borrowing may occasion agency and bankruptcy costs, choosing not to be a corporation involves restrictions on ownership and marketability that make equity less liquid and diversifiable. Thus, we would still expect the choice between corporate and non-corporate status to reflect a trade-off of tax and non-tax factors. But the types of firms choosing not to incorporate might differ from those choosing high debt–equity ratios within the corporate sector. The most obvious distinction relates to size, as larger entities that might have ready access to borrowed funds are likely to find the restrictions on non-corporate ownership structure very costly.

Finally, the relevant individual tax rates may differ when considering debt ownership and non-corporate status as alternatives to corporate equity ownership. While a large share of corporate debt is held by tax-exempt institutions, these institutions are effectively excluded from holding non-corporate equity in the United States by the imposition of unrelated business income taxes (UBIT). Thus, tax reforms might have different predicted effects on debt–equity and corporate–non-corporate choices. For example, the Tax Reform Act of 1986 sharply reduced the individual tax rate, but also reduced the corporate tax rate. Thus, tax-exempt institutions would have been pushed in the direction of holding equity versus debt (see expression 3.10), while high-bracket investors would have been pushed toward debt. However, only the latter group would be relevant when considering the corporate–non-corporate ownership choice, leading to a clearer prediction favoring a shift out of corporate form.

Considering the choice of organizational form adds a dimension to the measurement of corporate tax distortions. The classic analysis by Harberger (1966), empirically refined by Shoven (1976), treats the corporate and non-corporate sectors as distinct and exogenous, with distortions resulting from a misallocation of capital and labor between the two sectors. But the choice of organizational form introduces another potential

[19] Of course, a business seeking to constitute a portion of its operations in corporate form and another portion in non-corporate form can do so by spinning off one of the units as a separate entity. But this involves considerably higher transaction costs than an adjustment of the firm's debt level.

tax distortion into the picture. One cannot simply count distortions and conclude that this will make things worse, of course, for one distortion may mitigate another, in classic second-best fashion. For example, the favorable tax treatment of debt not only distorts the debt–equity decision, but also reduces the corporate cost of capital and may lessen the distortion of the choice between corporate and non-corporate status. Similarly, being able to adopt the tax-favored non-corporate form may lessen the capital allocation distortions associated with the corporate tax. It is a question to be resolved empirically how the choice of organizational form contributes to the distortions of the corporate tax. To date, though, most examinations have considered this distortion in isolation, rather than in conjunction with other distortions.

Perhaps the most straightforward method of estimating the distortions arising from choice of organizational form is first to estimate the sensitivity of this choice to variations in tax and non-tax factors, and then to apply these estimates in a deadweight loss calculation. MacKie-Mason and Gordon (1997) and Goolsbee (1998) take this approach by considering the share of aggregate assets in corporate and non-corporate form for the periods 1959–1986 and 1900–1939, respectively. Both studies find the corporate–non-corporate tax differential to have a significant but relatively small impact on the share of assets in corporate form[20]. As a result of the small measured effects on behavior, each paper's estimate of the deadweight loss arising purely from the corporate–non-corporate distortion is also relatively small, respectively 16 percent of business tax revenue and 5–10 percent of corporate tax revenue. These estimates stand in marked contrast to those derived by Gravelle and Kotlikoff (1989) using a calibrated simulation model based on a particular specification of the non-tax differences between corporate and non-corporate enterprises.

Further evidence of the role of tax factors in affecting organizational form comes from the period around the Tax Reform Act of 1986, after which elections of S corporation status surged almost immediately [Gordon and MacKie-Mason (1990)], particularly for profitable firms [Carroll and Joulfaian (1997)]. While not the primary focus of his comparison of the financial policies of master-limited partnerships and corporations in the oil and gas industry, Gentry (1994) does find that the choice between these two forms relates to certain tax and non-tax factors in predicted ways. For example, leverage appears to substitute for opting out of the corporate sector, and riskier firms are also less likely to choose partnership form.

4.2. Mergers and acquisitions

Mergers and acquisitions occur continually in a dynamic corporate environment, the ebbs and flows of activity being attributable to many factors, including the pace of

[20] MacKie-Mason and Gordon find further confirmation of the role of tax factors in the opposite responses of firms with losses and firms with gains, and the responses over time to changes in tax rules other than simple tax-rate changes.

technological change and the tone of government anti-trust policy. In response to a sharp surge in US merger activity in the 1980s, though, many attempts at explanation centered on the role of tax incentives.

Identifying the potential tax benefits of mergers and acquisitions confronts two significant obstacles at the outset. First, the tax law governing these transactions is complex. There are many different types of transactions within the general category, and the tax treatment of corporations and their shareholders varies by transaction type. Scholes and Wolfson (1992) provide a good discussion of the types of transactions and their tax consequences, but a brief summary is useful here. At the corporate level, the main distinction is whether the acquired company's tax attributes are carried over by the acquirer, or whether the target is treated as having been liquidated, in which case there are both immediate tax consequences and an establishment of new tax attributes. The parallel issue at the shareholder level is whether those who tender their shares in the target company are treated as having closed a position, or whether the tax basis in tendered shares carries over to the shares in the parent company that are received as payment. In order to qualify for tax-free treatment at the shareholder level (through a tax-free reorganization), the means of payment must be an exchange of stock, and the corporate attributes of the target must be carried over. Otherwise, for transactions that are taxable at the shareholder level, the corporate tax treatment is at the option of the acquiring company.

A second problem encountered in identifying the potential tax benefits of mergers and acquisitions is that benefits associated with mergers and acquisitions generally may be obtained through alternative transactions, though not necessarily as easily or at the same cost. Thus, the incremental tax benefits to merger and acquisition activity may be smaller than they might first appear to be. Indeed, Gilson et al. (1988) go through these alternatives and argue that the theoretical case for tax-induced merger activity is weak. Still, if firms are found to respond to the apparent tax incentives to merge, this suggests that they do not view the alternative means of obtaining tax benefits as perfect substitutes. Thus, the response of firms and markets to the tax incentives to merge remains an open question for empirical investigation.

4.2.1. Potential corporate tax benefits of mergers and acquisitions

There are three types of potential corporate tax benefits from a merger or acquisition: increased utilization of tax-loss and tax-credit carry-forwards, increased depreciation deductions obtained by stepping up the basis of assets, and increased interest deductions associated with an increase in the debt–equity ratio of the combined enterprise.

If a fully taxable firm acquires a firm with tax-loss and/or tax-credit carry-forwards, it may increase the value of these tax benefits by offsetting them against its own taxable income. The extent to which this increased utilization represents an incentive to merge depends on the available alternatives to using the benefits. As discussed above, a company can reduce the extent of its tax exhaustion by reducing its borrowing or by

leasing assets, but neither of these is a costless transaction; otherwise, we would not observe such a significant incidence of tax exhaustion in the first place.

An acquisition of a firm's unused tax deductions and credits is possible only if that firm's tax benefits are carried over by the acquirer. If, instead, the target is treated as having been liquidated, its tax books are closed, any unused tax shields disappear, any final corporate-level capital-gains taxes are due, and its assets are treated as if purchased directly by the acquirer. This has offsetting tax consequences. On the one hand, there may be taxes immediately payable. On the other hand, the present value of depreciation deductions may be substantially increased by an increase in asset basis and the ability to depreciate assets anew. In general such a trade (an immediate tax on basis step-up in exchange for the depreciation of this basis over time) might seem unattractive, but its appeal may be enhanced by a number of factors. First, the corporate capital-gains tax rate is below the rate at which depreciation allowances are deducted. Second, the new depreciation schedule might be more attractive than that being used by the target, because of changes in law. Third, there may be circumstances under which the initial capital-gains tax liability is forgiven when a corporation is liquidated. This was the case in the United States under the so-called General Utilities doctrine, until its repeal by the Tax Reform Act of 1986. This provision, while it existed, made the transfer of assets through a corporate acquisition more attractive than direct purchases of existing assets.

Perhaps the most problematic of the apparent corporate tax benefits from a merger or acquisition is increased interest deductions. This benefit looms large in perception, particularly with reference to leveraged buyouts in which a significant fraction of the cost of an acquisition is financed by newly issued debt. But in what sense is this a tax benefit connected to the acquisition, if the target company could have done the borrowing itself? There are two potential responses to this critique. First, a merger may pool the idiosyncratic risks of individual firms, reducing the non-tax borrowing costs for the combined entity. Second, existing management might be reluctant to borrow up to the value-maximizing level. In each case, an acquisition enhances the tax benefit of borrowing.

4.2.2. Potential shareholder tax benefits of mergers and acquisitions

In nontaxable stock transactions, there are no immediate tax consequences for shareholders, but there may still be implicit tax benefits. Tendering shareholders typically receive shares in a larger, more diversified enterprise, a process that can result in a more balanced portfolio without the capital-gains taxes usually attendant upon such rebalancing. In taxable cash transactions, the acquiring firm distributes cash out of corporate form at capital-gains rates, thus effecting share repurchases on a larger scale than those in which companies typically engage. As in the case of interest deductions, the tax benefit hinges on the acquisition facilitating the repurchase of shares.

4.2.3. Evidence on the role of taxes in mergers and acquisitions

Evidence concerning the impact of tax incentives on mergers and acquisitions may be adduced from patterns of merger activity and the market valuation of merger announcements. Each type of evidence provides at least some support for the argument that tax incentives affect mergers and acquisitions. Auerbach and Reishus (1988) considered a sample of mergers and acquisitions that took place during the period 1968–1983. They estimated that potential corporate tax benefits were significant in a number of cases, but also found that these benefits were not noticeably different from those that would have arisen from a matched sample of randomly chosen "pseudo-mergers". Using a "marriage model", they estimated that few of the apparent tax benefits affected whether a firm merged or the company with which it merged, the main exception being the tax status of the *acquiring* company. As the ability to offset unused tax shields against the new partner's taxable income is a benefit that applies symmetrically, tax-exhausted companies may be attractive targets, but apparently are even more likely to be energetic suitors.

Whatever the tax benefits to merging during the period studied by Auerbach and Reishus, things changed with the Tax Reform Act of 1986, due to the repeal of the General Utilities doctrine as well as newly imposed limits on the transfer of the tax benefits. These changes, in conjunction with the increased tax rate on the capital gains of individual shareholders, provided an incentive to time mergers to occur before January 1, 1987. Indeed, there was a strong surge in merger and acquisition activity, as measured by firm value, during the last quarter of 1986 [Scholes and Wolfson (1990)]. This does not necessarily contradict the finding of Auerbach and Reishus, for it is possible for tax factors to matter relatively little in determining whether transactions occur at all but enough to affect the timing of transactions within a short window. In fact, although the aggregate value of acquisitions did fall between 1986 and 1987, it surged again in 1988 and 1989 [Auerbach and Slemrod (1997)].

Evidence that tax factors play some role in influencing mergers and acquisitions – or at least that they *should* play some role, if managers strive to increase shareholder value – also comes from market responses to the announcement of acquisitions. In an event study covering the period 1970–1985, Hayn (1989) found the tax attributes of target firms to be significant in explaining the abnormal returns to shareholders in both target and acquiring firms, with loss and credit carry-forwards mattering for tax-free reorganizations and basis step-up mattering for taxable acquisitions. In an analysis of 76 management buyouts during the period 1980–1986, Kaplan (1989) suggested that a significant fraction of the buyout premiums, ranging from 21 percent to 143 percent, could be justified by tax benefits, depending on the imputed valuation of incremental interest deductions.

5. Taxes and financial innovation

As noted in the introduction to this chapter, financial policy decisions are typically measured in terms of observable categories such as debt and dividends, but fun-

damental financial decisions relate to the allocation of state-contingent claims. The perspective taken above generally assumes some given relationship between these nominal categories and underlying claims, so that the firm's choice, say, between debt and equity involves a trade-off between tax benefits and a different allocation of commitments across states of nature.

However, there may be flexibility in the correspondence between formal categories and underlying claims, and firms will seek to widen categories that are tax-preferred. In these instances, there is another type of trade-off, as attempts to extend favorable tax treatment to a wider class of financial claims may involve offsetting costs. Sometimes, the true social costs may be relatively trivial (e.g., the time of a good tax advisor) unless the government steps in to impose costs in the form of tax penalties or other legal sanctions. The benefits of the government's doing so are not always clear, though, as avoiding taxes does typically reduce distortions. For example, firms that successfully characterize as debt claims that possess equity-like characteristics can reduce their cost of capital, reducing the tax wedges facing corporate investment and the decision to operate as a corporation. But if this outcome represents an improvement to overall economic welfare, then why does the tax system attempt to distinguish between these two types of claims in the first place? Unless political inertia is such that improvements in social welfare must proceed through such "do-it-yourself" tax cuts, it is not clear what the optimal policy response is to such financial innovation.

Scholes and Wolfson (1992, Chapter 20) provide a cogent discussion of the difficulties tax authorities face in identifying financial innovation and the dilemma of what to do about it. They also provide many illustrations of such transactions. [Also see the discussion by Bulow et al. (1990)]. Not all of the choices available to firms involve the creation of exotic new combinations of claims, or "synthetic" assets. For example, a convertible bond is equivalent to a combination of a warrant and a straight bond, yet taxed differently. But the scope for tax arbitrage has continually widened, with new conceptions of how to break down and repackage contingent claims continually appearing. An illustration is the "unbundled stock units" briefly considered in the late 1980s, which would have divided equity claims into pieces to allow the dividend-paying portion to be treated as debt. Had such financial instruments taken root (the IRS having played a role in their not doing so), their large-scale use could have largely eliminated the corporate income tax.

The nature of tax arbitrage goes beyond the characterization of financial instruments. Even if the tax treatment of a particular transaction for a given taxpayer is clear, it may be possible and advantageous for one taxpayer to shift income to a related taxpayer subject to more favorable rules, as in the case of a domestic corporation and its foreign subsidiary[21]. Unrelated firms (or individuals) can engage in tax arbitrage by exchanging income and/or deductions, following the principle of comparative advantage based on relative tax treatment. This type of behavior is exemplified by

[21] See this Handbook's chapter by Gordon and Hines for further discussion (Volume 4, forthcoming).

the leasing transactions between taxable and nontaxable firms discussed above, but certainly not limited to such transactions.

As financial innovation can be expected to occur, applying increasingly to international as well as domestic transactions, perhaps the clearest conclusion one can draw is of the declining viability of tax systems attempting to enforce tax provisions that treat similar transactions inconsistently. Though the responsiveness of taxpayers to tax arbitrage opportunities may increase, the distortions of underlying financial policy may actually decline as a result. But the new sources of revenue needed to replace the funds lost to financial innovation will undoubtedly have distortions of their own.

References

Altshuler, R., and A.J. Auerbach (1990), "The significance of tax law asymmetries: an empirical investigation", Quarterly Journal of Economics 105:61–86.

American Law Institute (1982), Federal Income Tax Project, C: Proposals on Corporate Acquisitions and Distributions and Reporter's Study on Corporate Distributions (American Law Institute, Philadelphia).

Asquith, P.R., and D.W. Mullins (1986), "Equity issues and offering dilution", Journal of Financial Economics 15:61–89.

Auerbach, A.J. (1979a), "Share valuation and corporate equity policy", Journal of Public Economics 11:291–305.

Auerbach, A.J. (1979b), "Wealth maximization and the cost of capital", Quarterly Journal of Economics 93:433–446.

Auerbach, A.J. (1981), "Tax integration and the new view of the corporate tax: a 1980s perspective", in: Proceedings of the National Tax Association (National Tax Association, Columbus, OH) pp. 21–27.

Auerbach, A.J. (1983a), "Taxation, corporate financial policy, and the cost of capital", Journal of Economic Literature 21:905–40.

Auerbach, A.J. (1983b), "Stockholder tax rates and firm attributes", Journal of Public Economics 21:107–27.

Auerbach, A.J. (1984), "Taxes, firm financial policy and the cost of capital: an empirical analysis", Journal of Public Economics 23:27–57.

Auerbach, A.J. (1985), "Real determinants of corporate leverage", in: B. Friedman, ed., Corporate Capital Structures in the United States (University of Chicago Press, Chicago) pp. 301–22.

Auerbach, A.J. (1990), "Debt, equity and the taxation of corporate cash flows", in: J. Shoven and J. Waldfogel, eds., Taxes and Corporate Restructuring (Brookings, Washington, DC) pp. 91–126.

Auerbach, A.J., and K.A. Hassett (2000), "On the marginal source of investment funds", Journal of Public Economics, forthcoming.

Auerbach, A.J., and M.A. King (1983), "Taxation, portfolio choice, and debt–equity ratios: a general equilibrium model", Quarterly Journal of Economics 98:587–609.

Auerbach, A.J., and L.J. Kotlikoff (1987), Dynamic Fiscal Policy (Cambridge University Press, Cambridge, UK).

Auerbach, A.J., and J.M. Poterba (1987), "Tax-loss carryforwards and corporate tax incentives", in: M. Feldstein, ed., The Effects of Taxation on Capital Accumulation (University of Chicago Press, Chicago) pp. 305–338.

Auerbach, A.J., and D. Reishus (1988), "The effects of taxation on the merger decision", in: A. Auerbach, ed., Corporate Takeovers: Causes and Consequences (University of Chicago Press, Chicago) pp. 157–183.

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

Bagwell, L.S., and J.B. Shoven (1989), "Cash distributions to shareholders", Journal of Economic Perspectives 3(Summer):129–140.
Barclay, M., and C.W. Smith Jr (1988), "Corporate payout policy: cash dividends versus open-market repurchases", Journal of Financial Economics 22:61–82.
Barclay, M., and C.W. Smith Jr (1995), "The priority structure of corporate liabilities", Journal of Finance 50:899–917.
Bernheim, B.D. (1991), "Tax policy and the dividend puzzle", Rand Journal of Economics 22:455–476.
Bernheim, B.D., and A. Wantz (1995), "A tax-based test of the dividend-signaling hypothesis", American Economic Review 85:532–551.
Black, F., and M.S. Scholes (1973), "The pricing of options and corporate liabilities", Journal of Political Economy 81:637–659.
Boadway, R., and N. Bruce (1992), "Problems with integrating corporate and personal income taxes in an open economy", Journal of Public Economics 48:39–66.
Bond, S., and C. Meghir (1994), "Dynamic investment models and the firm's financial policy", Review of Economic Studies 61:197–222.
Bradford, D. (1981), "The incidence and allocation effects of a tax on corporate distributions", Journal of Public Economics 15:1–22.
Bradley, M., G.A. Jarrell and E.H. Kim (1984), "On the existence of an optimal capital structure: theory and evidence", Journal of Finance 39:857–878.
Brealey, R.A., and C.M. Young (1980), "Debt, taxes and leasing – a note", Journal of Finance 35:1245–1250.
Brennan, M., and A.V. Thakor (1990), "Shareholder preferences and dividend policy", Journal of Finance 45:993–1018.
Bulow, J.I., L.H. Summers and V.P. Summers (1990), "Distinguishing debt from equity in the junk bond era", in: J. Shoven and J. Waldfogel, eds., Taxes and Corporate Restructuring (Brookings, Washington, DC) pp. 135–166.
Carroll, R., and D. Joulfaian (1997), "Taxes and corporate choice of organizational form", OTA Paper 73 (U.S. Department of the Treasury) October.
Cummins, J.G., K.A. Hassett and R.G. Hubbard (1994), "A reconsideration of investment behavior using tax reforms as natural experiments", Brookings Papers on Economic Activity 25:1–59.
DeAngelo, H., and R.W. Masulis (1980), "Optimal capital structure under corporate and personal taxation", Journal of Financial Economics 8:3–29.
Devereux, M.P., and H. Freeman (1995), "The impact of tax on foreign direct investment: empirical evidence and the implications for tax integration schemes", International Tax and Public Finance 2:85–106.
Edwards, J.S.S., and M.J. Keen (1984), "Wealth maximization and the cost of capital: a comment", Quarterly Journal of Economics 99:211–214.
Edwards, J.S.S., and C.P. Mayer (1991), "Leasing, taxes, and the cost of capital", Journal of Public Economics 44:173–197.
Fama, E.F., and K.R. French (1998), "Taxes, financing decisions, and firm value", Journal of Finance 53:819–843.
Fama, E.F., and K.R. French (2000), ay?" Working Paper 509 (CRSP, Chicago).
Fazzari, S.M., R.G. Hubbard and B.C. Petersen (1988), "Financing constraints and corporate investment", Brookings Papers on Economic Activity 19:141–195.
Froot, K.A., and J.R. Hines Jr (1995), "Interest allocation rules, financing patterns, and the operations of U.S. multinationals", in: M. Feldstein, J. Hines and G. Hubbard, eds., The Effects of Taxation on Multinational Corporations (University of Chicago Press, Chicago) pp. 277–307.
Gentry, W. (1994), "Taxes, financial decisions and organizational form: evidence from publicly traded partnerships", Journal of Public Economics 53:223–244.
Gilson, R.J., M.S. Scholes and M.A. Wolfson (1988), "Taxation and the dynamics of corporate control: the uncertain case for tax-motivated transactions", in: J. Coffee, L. Lowenstein and S. Rose-Ackerman,

eds., Knights, Raiders, and Targets: the Impact of the Hostile Takeover (Oxford University Press, New York) pp. 271–299.
Givoly, D., C. Hayn, A.R. Ofer and O. Sarig (1992), "Taxes and capital structure: evidence from firms' response to the Tax Reform Act of 1986", Review of Financial Studies 5:331–355.
Goolsbee, A. (1998), "Taxes, organizational form, and the deadweight loss of the corporate income tax", Journal of Public Economics 69:143–152.
Gordon, R.H., and J.K. MacKie-Mason (1990), "Effects of the Tax Reform Act of 1986 on corporate financial policy and organizational form", in: J. Slemrod, ed., Do Taxes Matter? The Impact of the Tax Reform Act of 1986 (MIT Press, Cambridge) pp. 91–131.
Graham, J.R. (1996), "Debt and the marginal tax rate", Journal of Financial Economics 41:41–73.
Graham, J.R. (1999), "Do personal taxes affect corporate financing decisions?" Journal of Public Economics 73:147–185.
Graham, J.R., M.L. Lemmon and J.S. Schallheim (1998), "Debt, leases, taxes, and the endogeneity of corporate tax status", Journal of Finance 53:131–162.
Gravelle, J.G., and L.J. Kotlikoff (1989), "The incidence and efficiency costs of corporate taxation when corporate and non-corporate firms produce the same good", Journal of Political Economy 97:749–780.
Harberger, A.C. (1966), "Efficiency effects of taxes on income from capital", in: M. Krzyzaniak, ed., Effects of the Corporation Income Tax (Wayne State University Press, Detroit) pp. 107–117.
Hayn, C. (1989), "Tax attributes as determinants of shareholder gains in corporate acquisitions", Journal of Financial Economics 23:121–153.
Jensen, M. (1986), "Agency costs of free cash flow, corporate finance, and takeovers", American Economic Review 76:323–329.
Judd, K.L., and B.C. Petersen (1986), "Dynamic limit pricing and internal finance", Journal of Economic Theory 39:268–299.
Kanniainen, V., and J. Södersten (1994), "Costs of monitoring and corporate taxation", Journal of Public Economics 55:307–321.
Kanniainen, V., and J. Södersten (1995), "The importance of reporting conventions for the theory of corporate taxation", Journal of Public Economics 57, 417–430.
Kaplan, S. (1989), "Management buyouts: evidence on taxes as a source of value", Journal of Finance 44:611–632.
King, M.A. (1974), "Taxation and the cost of capital", Review of Economic Studies 41:21–35.
King, M.A. (1977), Public Policy and the Corporation (Chapman and Hall, London).
King, M.A., and D. Fullerton, eds (1984), The Taxation of Income from Capital (University of Chicago Press, Chicago).
Lyon, A.B. (1990), "Investment incentives under the alternative minimum tax", National Tax Journal 43:451–465.
MacKie-Mason, J.K. (1990), "Do taxes affect corporate financing decisions?" Journal of Finance 45:1471–1493.
MacKie-Mason, J.K., and R.H. Gordon (1997), "How much do taxes discourage incorporation?" Journal of Finance 52:477–505.
Miller, M.H. (1977), "Debt and taxes", Journal of Finance 32:261–275.
Miller, M.H., and F. Modigliani (1961), "Dividend policy, growth, and the valuation of shares", Journal of Business 34:411–433.
Modigliani, F., and M.H. Miller (1958), "The cost of capital, corporation finance, and the theory of investment", American Economic Review 48:261–297.
Myers, S.C. (1977), "Determinants of corporate borrowing", Journal of Financial Economics 5:147–175.
Myers, S.C., and N. Majluf (1984), "Corporate financing and investment decisions when firms have information that investors do not have", Journal of Financial Economics 13:187–221.
Myers, S.C., D.A. Dill and A.J. Bautista (1976), "Valuation of financial lease contracts", Journal of Finance 31:799–819.

Poterba, J.M., and L.H. Summers (1985), "The economic effects of dividend taxation", in: E. Altman and M. Subrahmanyam, eds., Recent Advances in Corporate Finance (Richard D. Irwin, Homewood, IL) pp. 227–284.

Rajan, R.G., and L. Zingales (1995), "What do we know about capital structure? Some evidence from international data", Journal of Finance 50:1421–1460.

Samuelson, P.A. (1964), "Tax deductibility of economic depreciation to insure invariant valuations", Journal of Political Economy 72:604–606.

Scholes, M.S., and M.A. Wolfson (1990), "The effects of changes in tax laws on corporate reorganization activity", Journal of Business 63:S141–S164.

Scholes, M.S., and M.A. Wolfson (1992), Taxes and Business Strategy (Prentice-Hall, Englewood Cliffs, NJ).

Scholes, M.S., G.P. Wilson and M.A. Wolfson (1990), "Tax planning, regulatory capital planning, and financial reporting strategy for commercial banks", Review of Financial Studies 3:625–650.

Sharpe, S.A., and H.H. Nguyen (1995), "Capital market imperfections and the incentive to lease", Journal of Financial Economics 39:271–294.

Shoven, J.B. (1976), "The incidence and efficiency effects of taxes on income from capital", Journal of Political Economy 84:1261–1283.

Sinn, H.-W. (1987), Capital Income Taxation and Resource Allocation (North Holland, Amsterdam).

Sinn, H.-W. (1991a), "The vanishing Harberger triangle", Journal of Public Economics 45:271–300.

Sinn, H.-W. (1991b), "Taxation and the cost of capital: the 'old' view, the 'new' view, and another view", in: D. Bradford, ed., Tax Policy and the Economy, Vol. 5 (MIT Press, Cambridge, MA) pp. 25–54.

Sørensen, P.B. (1994), "Some old and new issues in the theory of corporate income taxation", Finanz Archiv 51:425–456.

Stiglitz, J.E. (1973), "Taxation, corporate financial policy and the cost of capital", Journal Public Economics 2:1–34.

Stiglitz, J.E., and A. Weiss (1981), "Credit rationing in markets with imperfect information", American Economic Review 71, 393–410.

Titman, S., and R. Wessels (1988), "The determinants of capital structure choice", Journal of Finance 43:1–19.

Chapter 20

TAX POLICY AND BUSINESS INVESTMENT*

KEVIN A. HASSETT

American Enterprise Institute, Washington, DC

R. GLENN HUBBARD

Graduate School of Business, Columbia University, New York, NY

Contents

* RGH acknowledges financial support from Harvard Business School and the American Enterprise Institute.

Handbook of Public Economics, Volume 3, Edited by A.J. Auerbach and M. Feldstein

Abstract

In this survey, we review research on tax policy and business investment with four objectives. First, we use a simple prototypical dynamic neoclassical investment model to derive and explain effects of taxation on business investment in the long run and short run. Second, we describe and evaluate empirical tests of neoclassical channels, and we conclude that recent empirical evidence is consistent with neoclassical intuition. Third, we explore qualifications to basic theoretical models and their empirical tests raised by recent research on irreversibility and capital-market imperfections. Finally, we evaluate arguments for and against using tax policy to influence the level or timing of investment.

While there is a consensus about the nature and magnitude of tax policy on investment demand considerable uncertainty remains regarding the structure of adjustment costs and the short-run dynamic effects of tax reforms. Consistent with our analysis of equilibrium investment outcomes, ascertaining the effects of tax policy on equilibrium investment requires additional research to examine responsiveness of interest rates, output, and the stock market to tax policy changes.

Keywords

tax policy, investment

JEL classification: H20, H25, H21

1. Introduction

Economists have long argued that significant reforms of personal and company taxation can have large effects on firms' investment decisions. At some level policymakers themselves have heeded this message. During the 1980s, for example, significant tax reforms were introduced in many countries [see Messere (1993)]. During the 1990s, continued discussion in the United States of corporate tax reform and of the desirability of switching from an income-based tax to a consumption-based tax system has centered on effects on investment and capital formation. Nevertheless, while extensive studies have examined the effects of tax parameters on the cost of or returns to investment, empirical evidence is mixed.

In this survey, we review research on tax policy and business investment with four objectives. First, we use a simple prototypical dynamic neoclassical investment model to derive and explain effects of taxation on business investment in the long run and short run. Second, we describe and evaluate empirical tests of neoclassical channels, and we conclude that recent empirical evidence is consistent with neoclassical intuition. Third, we explore qualifications to basic theoretical models and their empirical tests raised by recent research on irreversibility and capital-market imperfections. Finally, we evaluate arguments for and against using tax policy to influence the level or timing of investment.

Toward these ends, the review is organized as follows. Section 2 develops a simple equilibrium model to identify links between tax policy and fixed capital accumulation. In Sections 3 and 4, we review empirical evidence on the short-run and long-run responsiveness of investment to changes in tax parameters. Section 5 focuses on arguments supporting or opposing the use of tax policy to affect investments. In Section 6, we apply the lessons from recent research to an evaluation of lower inflation or a switch to consumption taxation. Section 7 concludes.

2. Tax policy, investment, and capital accumulation

We begin with a simple general equilibrium model of the economy to investigate effects of tax parameters on investment and the capital stock in steady state and in the short run[1]. In most of what follows, we consider only a real economy, though we describe more heuristically later in this review the effects of inflation on the capital stock.

2.1. Households

For simplicity, we focus on a model of a representative infinitely-lived consumer choosing consumption (C) and labor supply (L). Savings may be allocated between

[1] This treatment follows Summers (1981), Judd (1985), Sinn (1987) and Turnovsky (1995).

government bonds (B) and business equity (E); we abstract from business debt. Government consumption spending has no direct effect on households' utility. Letting ρ represent the rate of time preference, we assume the household maximizes

$$\int_0^\infty U(C,L)\,\mathrm{e}^{-\rho t}\,\mathrm{d}t, \quad \text{where} \quad U_C > 0,\ U_L < 0; \quad U_{CC}, U_{LL}, U_{CL} < 0 \tag{1}$$

subject to the budget constraint

$$\dot{B} + s\dot{E} + C = (1 - t_\mathrm{p})[wL + rB + D] - t_\mathrm{g}\dot{s}E + R \tag{2}$$

and initial conditions

$$B(0) = B_o \quad \text{and} \quad E(0) = E_o, \tag{3}$$

where s = price of equities relative to current output, t_p = ordinary income tax rate, w = real wage rate, r = real interest rate on government bonds, D = dividends, t_g = tax rate on capital gains, R = lump-sump tax rebate to consumers.

Consumers' optimality conditions are given by

$$U_C' = \lambda, \tag{4a}$$
$$U_L' = -w(1 - t_\mathrm{p})\,\lambda, \tag{4b}$$
$$r(1 - t_\mathrm{p}) = \beta, \tag{4c}$$
$$(1 - t_\mathrm{p})(D/sE) + (1 - t_\mathrm{g})(\dot{s}/s) = \beta, \tag{4d}$$
$$\rho - (\dot{\lambda}/\lambda) = \beta, \tag{4e}$$
$$\lim_{t \to \infty} \lambda B \mathrm{e}^{-\rho t} = \lim_{t \to \infty} \lambda E \mathrm{e}^{-\rho t} = 0. \tag{4f}$$

Equation (4a) defines the marginal utility of wealth, λ; Equation (4b) defines the equilibrium wage. The rate of return on consumption β equals the after-personal-tax real rate of interest (4c). The after-tax real return on equities, represented by the left-hand side of Equation (4d), equals the real after-tax return on bonds (that is, both equal β). Finally, the optimality equations include the transversality conditions (4f).

2.2. Firms

Models in the neoclassical tradition focus on the derived demand for capital by value-maximizing firms[2]. This intuition is typically transformed into models of investment

[2] For the exercise we consider here, we focus on investment decisions of domestic corporations. The same intuition may be applied to investment decisions by entrepreneurs [in which personal taxes play a larger role, as in Carroll, Holtz-Eakin, Rider and Rosen (2000)] and investment decisions by multinational corporations (in which residence-country and source-country taxes play a role).

by making assumptions about costs of changing the capital stock. For simplicity of exposition, we consider the decisions of a price-taking firm. Absent taxes, firms net cash flow (X) is given by

$$X = F(K,L) - wL - pI - \Psi(I,K), \tag{5}$$

where $F(K,L)$ is a well-behaved neoclassical production function; K is the capital stock; L is labor; p is the price of investment goods related to current output; I is investment; and Ψ is the function determining the cost of adjusting the capital stock[3]. In the absence of taxes, then, the marginal cost of newly installed capital is $p + \Psi_I(I,K)$.

To study investment tax policy, we add to the net cash flow expression (5) a corporate tax rate of t_c, an investment tax credit at rate ITC, and the present value of a dollar's worth of depreciation allowance, z [4]. With the additions, the marginal cost of newly installed capital is

$$p(1-\Gamma) + (1-t_C)\,\Psi_I(I,K),$$

where $\Gamma = \text{ITC} + t_c z$, so that we can rewrite Equation (5) as

$$X = (1-t_C)(F(K,L) - wL - \Psi_I(I,K)) - p(1-\Gamma)I. \tag{5'}$$

Following the "tax capitalization" view of the dividend decision [see King (1977), Auerbach (1979) and Bradford (1981)], we make dividends a residual in business decisions. That is, dividends are assumed to equal net cash flow after investment[5]. Retained earnings are the marginal source of funds for investment until they are exhausted; at that point, new equity issues are the marginal source of funds for investment.

The value of outstanding business equities V is

$$V = sE. \tag{6}$$

Differentiating Equation (6) with respect to time (t) and using Equations (4d) and (5') yields the following differential equation in V:

$$\dot{V} = \frac{\beta}{1-t_g} V - \frac{(1-t_p)}{(1-t_g)}[(1-t_C)\pi - I - \Psi(I,K)]. \tag{7}$$

[3] See the discussion in Eisner and Strotz (1963), Lucas (1967), Gould (1968), Hayashi (1982) and Cummins, Hassett and Hubbard (1994, 1996).

[4] For the sake of simplifying the discussion, we focus here on the US tax system. For a parallel analysis that employs a more general tax formulation nesting that of many countries, see King and Fullerton (1984) and Sinn (1987).

[5] An alternative "traditional view" of the dividend decision argues that shareholders value dividends and that new equity issues are the marginal source of funds for investment [see the review in Poterba and Summers (1985)].

where

$$\pi = F(K, L) - wL.$$

The coefficient of V represents the discount rate applied by corporations to after-corporate-tax cash flows. In the derivation we present here, this discount rate is not affected by the dividend tax (t_p); for a discussion of alternative links among dividend taxes, dividend payouts, and the cost of capital, see Auerbach's chapter 19 in this volume.

The firm chooses I, K and L to maximize the firm's individual value $V(0)$ (i.e., $max \int_0^\infty e^{-pt} x_t \, dt$).

Optimality conditions are given by

$$F_L(K, L) = w, \tag{8a}$$

$$\frac{(1 - t_p)}{(1 - t_g)} [p(1 - \Gamma) + (1 - t_c)\Psi_1(I, K)] = q, \tag{8b}$$

$$\frac{(1 - t_p)}{(1 - t_g)} \left[\frac{(1 - t_c) F_K(K, L)}{q} + \frac{\dot{q}}{q} + \frac{(qI - HK)}{qk} \right] = \frac{\beta}{(1 - t_g)}, \tag{8c}$$

where $H = (I/K) + \Psi(I/K, 1)$.

Conditions (8b) and (8c) correspond to the "q" and "user cost" terms frequently used in neoclassical investment models. Abstracting from personal taxes, Equation (8b) states that the firm should invest up to the point at which the tax-adjusted price of the capital good $[p(1 - \Gamma)]$ equals its net-of-adjustment-cost shadow value $[q - (1 - t_C)\, \Psi_1(I, K)]$; see Hayashi (1982) and Summers (1981). Equation (8c) defines the user cost of capital.

We can also use this set-up to link q to the value of equities. Following Hayashi (1982), using the linear homogeneity of the production function and maintaining an assumption of perfect competition:

$$q = \frac{\hat{V}}{K} = \frac{sE}{K}. \tag{9}$$

2.3. Government sector

The government sector acts according to its cash-flow budget constraint:

$$\dot{B} = G + rB - t_p(wL + RB + D) - t_g \dot{s}E - t_c \pi + R, \tag{10}$$

where G represents real government expenditure (which we assume below remains constant over time). If the government changes t_p, t_c or t_g, it adjusts the path of debt (B) and/or lump-sum rebates (R) to ensure that its intertemporal budget constraint is satisfied. Note that we simplify here by assuming that there is no net private debt.

2.4. Equilibrium

To obtain the economy's equilibrium, we must put together optimality conditions for households and firms, the government's budget constraint, and market-clearing conditions. This equilibrium is characterized by the static condition

$$U_C' = \lambda, \tag{4a'}$$

$$U_L' = -F_L(1 - t_p)\lambda, \tag{4b'}$$

$$I = \phi(q)K \quad \text{and} \quad \phi'(q) > 0, \tag{11}$$

where condition (11) represents the solution of Equation (8b) in a "Tobin's q" setup. We can use Equations (4a') and (4b') to solve for consumption and employment:

$$C = C(\lambda, K, t_p), \quad C_\lambda < 0,\ C_K < 0,\ C_{t_p} > 0, \tag{12a}$$

$$L = L(\lambda, K, t_p), \quad L_\lambda > 0,\ L_K > 0,\ L_{t_p} < 0. \tag{12b}$$

Next, we can use Equations (11), (12a) and (12b) to express I, C and L in terms of q, K and λ:

$$I = \dot{K} = \phi(q)K, \tag{13a}$$

$$\dot{q} = \frac{\beta}{1 - t_g}q - (1 - t_c)F_K(K, L(\lambda, K, t_p)) - \phi(q)q + H[\Phi(q)]. \tag{13b}$$

We can determine the rate of return on consumption, β, from the product–market equilibrium condition:

$$F(K, L(\lambda, K, t_p)) = c(\lambda, K, t_p) + H[\phi(q)]K + G, \tag{13c}$$

$$\dot{\lambda} = \lambda(\rho - \beta), \tag{4e'}$$

so that

$$\beta = \beta(\lambda, K, q, t_p, t_c, t_g). \tag{13d}$$

2.5. Steady-state effects of tax policy on the capital stock

Without going into details about the solution of the model [see, e.g., Turnovsky (1995)], we can describe the steady state of the system. The steady state is reached where $\dot{q} = \dot{\lambda} = \dot{k} = 0$.

Net investment equals zero in the steady state, so the long-run value of equities equals that of the capital stock less the capitalized value of the dividend tax, or

Table 1
Qualitative steady-state effects of tax changes

An increase in	Affects				
	Capital stock (K)	Output (Y)	Consumption (C)	q	Cost of capital
t_c	−	−	−	*no change*	+
t_p	−	−	−	−	0
t_g	−	−	−	+	+
Γ	+	+	+	−	−

$q^* = (1 - t_p)/(1 - t_g)$. The rate of time preference (ρ) and the rate of return on consumption (β) are equal. The steady-state values K^*, L^* and C^* are reflected in

$$U_L(C^*, L^*) + F_L(K^*, L^*)(1 - t_p)\, U_C(C^*, L^*) = 0, \tag{14a}$$

$$(1 - t_c) F_K(K^*, L^*) = \frac{\rho}{(1 - t_g)}, \tag{14b}$$

$$F_K(K^*, L^*) = C^* + G. \tag{14c}$$

Tax policy affects the steady-state solutions for K, L and C. The corporate tax rate (t_c), the personal income tax rate (t_p), and the capital gains tax rate (t_g) affect the required return on capital [see Equation (14b)] and the equilibrium capital–labor ratio. The personal income tax rate affects the supply of labor. Below we consider steady-state responses of the capital–labor ratio and other variables of interest to permanent changes in the tax parameters or in investment incentives.

Corporate tax rate (t_c). From Equation (14b), an increase in the corporate tax rate, all other things being equal, increases the pre-tax physical product of capital (despite an increase in the value of depreciation deductions) so the capital–labor ratio falls, reducing consumption and output. Because $q = (1 - t_p)/(1 - t_g)$, steady-state stock prices are unaffected by the tax change, and the cost of capital is unaffected.

Personal income tax rate (t_p). An increase in the personal income tax raises the interest rate (4c) and the required return on capital (noting that $\rho = \beta$ in the steady state), reducing the capital–labor ratio (14b), consumption, and output. Steady-state stock prices fall because of the assumption of dividend-tax capitalization. In many countries there is a separate tax rate for income from dividends, but we simplify here by assuming that dividend is income.

Capital-gains tax rate (t_g). An increase in the capital-gains tax raises the required return on capital, reducing the capital–labor ratio, consumption, and output. For a shareholder to be indifferent between receiving a dividend and having the funds reinvested as retained earnings, q must rise to offset the higher capital-gains tax.

Investment incentives (Γ). An increase in investment incentives Γ reduces q (8b), and increases the capital–labor ratio, investment, output, and consumption. Normally, this depends on the corporate tax rate.

Table 1 summarizes these effects of permanent changes in t_c, t_p, t_g and Γ on the capital stock, output, consumption, q, and the cost of capital.

2.6. *Dynamic effects of tax policy shocks*

To study short-run effects of tax policy on investment, we assume dividend-tax capitalization and ignore depreciation; we examine the phase diagram for the $\dot{K} = 0$ and $\dot{q} = 0$ loci, see Figure 1, which are derived from Equations (13a,b)[6]. In each case, we consider the short-run effect of an unanticipated permanent tax change on the capital stock, output, consumption, q, and the cost of capital.

Corporate tax rate (t_c). An increase in the corporate tax shifts the $\dot{q} = 0$ locus down, while the $\dot{K} = 0$ locus is unaffected; the equilibrium capital stock falls from K_0 to K^*, as Figure 1a shows. In the short run, q falls, then gradually rises as capital is decumulated.

Personal tax rate (t_p). An increase in the personal income tax rate reduces q in the short run; investment falls, and q and K trace this locus to the new steady state ($K_1^* < K_0^*$) as in Figure 1b in which consumption is higher and output lower.

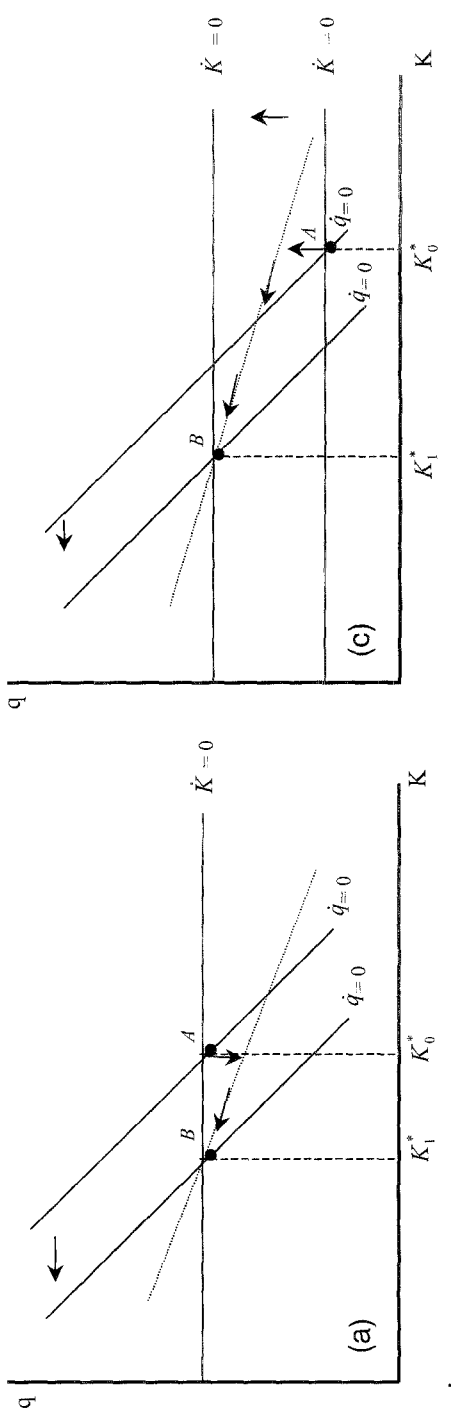
Capital-gains tax rate (t_g). An increase in the capital-gains tax rate raises the long-run cost of capital, reducing the long-run capital stock, producing similar dynamics to those accompanying an increase in the personal tax rate, t_p. One difference is that, ultimately, q returns to a value higher than at which it began, as Figure 1c shows.

Investment incentives (Γ). A permanent increase in the generosity of investment incentives reduces q and increases the long-run capital stock over time, as Figure 1d shows.

Temporary tax changes. With a temporary tax increase in t_p, the drop in the capital stock is not so pronounced as in the case of a permanent tax change. In addition, q declines less initially and actually overshoots its new equilibrium value to provide firms the incentive to restore their now-lower capital stock to its original level, as Figure 2 shows for the case of a temporary increase in the personal income tax rate, t_p. Similar intuition prevails for temporary changes in t_c and t_g. While a temporary investment tax credit reduces q by less than a permanent credit the temporary credit leads to a short-run boom in investment as firms attempt to increase investment before the credit expires.

Table 2 summarizes these short-run effects of changes in t_c, t_p, t_g and Γ on I, q and the cost of capital.

[6] This discussion follows Abel (1990).



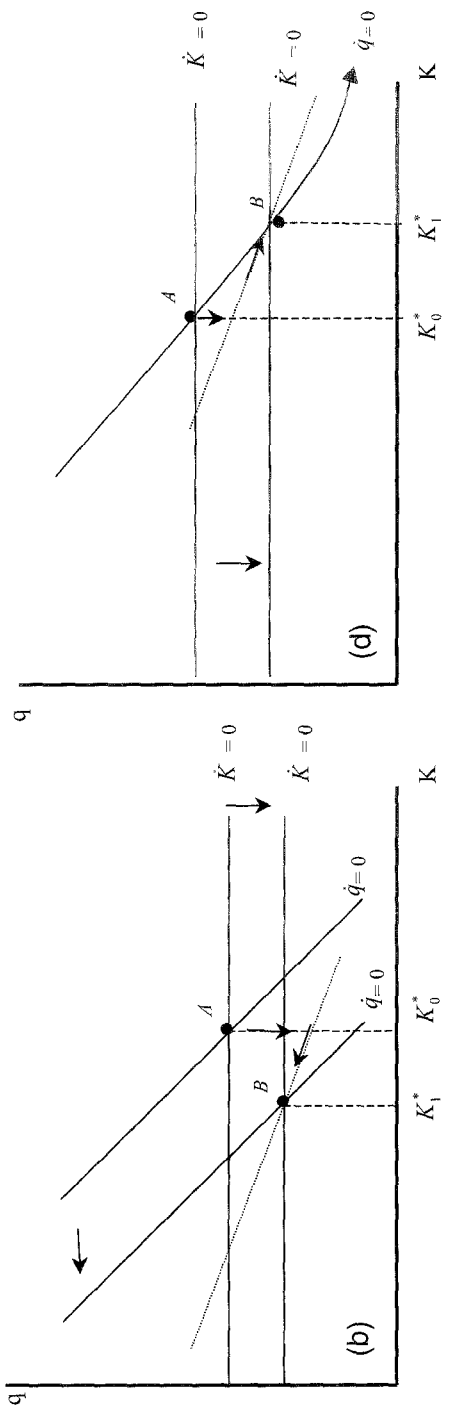


Fig. 1. Dynamic effects of tax policy changes: (a) unanticipated increase in t_c; (b) unanticipated increase in t_p; (c) unanticipated increase in t_g; (d) unanticipated increase in Γ. The dotted lines represent the saddle path.

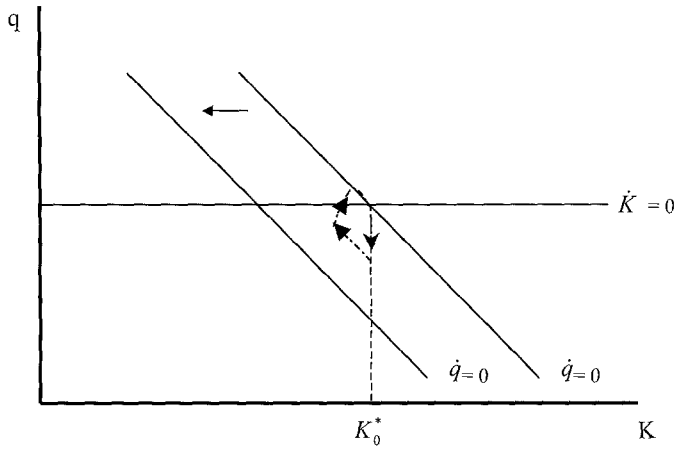


Fig. 2. Effects of a temporary increase in t_p.

Table 2
Qualitative short-run effects of tax changes

An increase in	Affects		
	I	q	Cost of capital
t_c	–	–	–
t_p	–	–	–
t_g	–	–	?
Γ	+	–	–

2.7. *Irreversibility and uncertainty*

The simple derivation thus far abstracts from the possibility that capital cannot be resold frictionlessly after being installed. With nonconvex costs of adjusting the capital stock, the analysis of investment dynamics differs from the case of convex costs of adjusting the capital stock, in part because the decision to invest exhausts an option of delay, the value of which introduces a range in which investment is less responsive to changes in neoclassical fundamentals [see, e.g., the excellent review of studies in Dixit and Pindyck (1994) and Caballero (1999)]. As Caballero, Engel and Haltiwanger (1995) point out, however, the steady-state implications of these models are often similar to those derived above in the context of neoclassical models with convex adjustment costs.

We can explore effects of irreversibility in our analytical framework by considering the case in which adjustment costs are not quadratic (the typical formulation of convex adjustments), as would be the case if firms face higher marginal costs of decreasing the capital stock than increasing the capital stock. In this case, good news

about neoclassical fundamentals causes the capital stock to increase relatively quickly (and q is not above its long-run value for an extended period of time), while bad news about fundamentals leads to relatively slow capital decumulation (and q is below its long-run value for a long period of time).

Studies of irreversibility generally examine consequences of uncertainty about fundamentals as well. Uncertainty about fundamentals affects expected future fundamentals and investment when adjustment costs are not quadratic or when profits are not linear in K (as would be the case, for example, in the absence of constant returns to scale and perfect competition).

A simple example makes the point. Suppose there is uncertainty about the intercept of the profit function [following Romer (1996)]. In an upcoming election, a key proposal is a major cut in corporate-profits taxes, which has a 50 percent chance of being adopted. The case of convex adjustment costs is illustrated in Figure 3a. Once the vote takes place, the expected capital gain is zero, and K and q follow the appropriate saddle path to the "proposal adopted" or "proposal rejected" long-run equilibrium. Given the 50–50 probabilities, q is midway between the points on the two saddle paths at the time of the vote (point A). A possible path to equilibrium following the vote is depicted in the chart.

With nonconvex costs of adjusting the capital stock (Figure 3b), q is also midway between the two saddle paths. When the proposal is announced, q jumps to the point where the dynamics of q and K move them to the relevant long-run equilibrium. However, the initial jump in q is not as great as in the case of quadratic adjustment costs. This is because it is relatively costly to reduce capital stocks accumulated before the election, reducing both the pre-election value of capital and investment. It is this effect that is commonly referred to as an option value: the firm retains the option of keeping its capital stock low; higher investment exercises this option.

This example highlights the important role that the adjustment cost function plays in determining the dynamic effects of policy. Researchers often also consider cases where the marginal cost of the first unit of investment is strictly positive, or where there are fixed costs to undertaking any nonzero investment. Both adaptations create often broad ranges of q where it is optimal to have a zero investment.

The simple neoclassical model of capital accumulation we summarized above suggest four challenges for empirical researchers attempting to estimate the sensitivity of fixed investment or the capital stock to neoclassical fundamentals (including tax rates and investment incentives). First, focusing on the long-run relationship between investment and neoclassical fundamentals, can one isolate shocks to a fundamental independent of shocks to other variables in firms' environment? Second, can one identify long-run tax changes to investigate their effects on the level and timing of investment? Third, is it possible to mitigate potential measurement error in tax and non-tax components of the marginal profitability of capital or the user cost of capital? Finally, how well do the simple model's predictions based on convex costs of adjusting the capital stock describe investment dynamics?

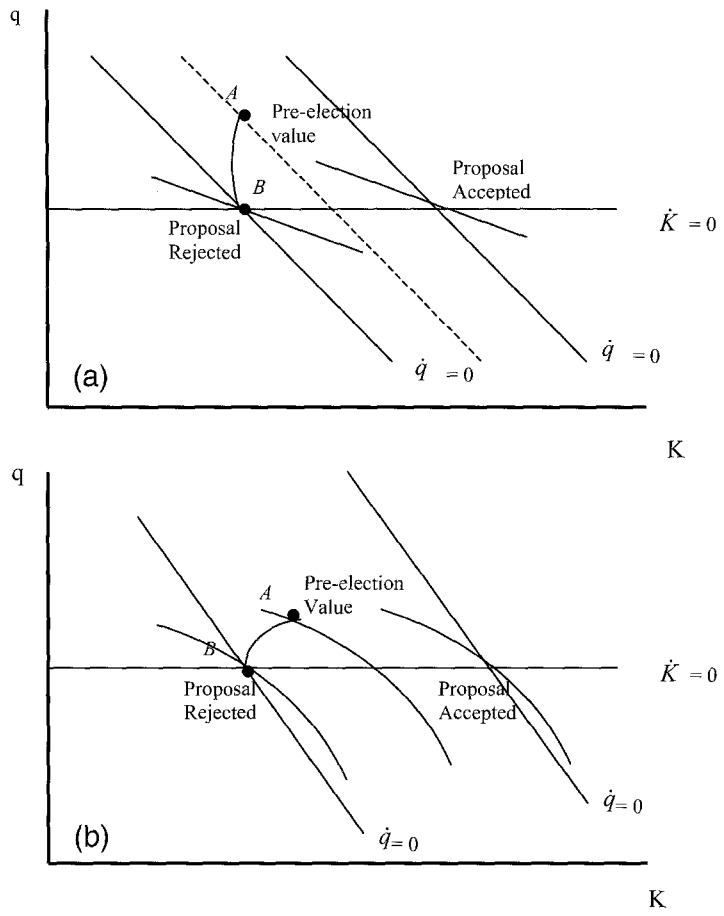


Fig. 3. Investment when future tax policy is uncertain: (a) quadratic adjustment costs; (b) asymmetric adjustment costs.

With these challenges in mind, we turn to empirical research. Perhaps surprisingly given the strong predictions of simple neoclassical models, much of the empirical debate until the past decade has centered on whether personal and business taxation have *any* effect on firms' investment decisions. More recently, debates have become more subtle, though no less vigorous.

3. Moving from analytical to empirical analysis of investment

Empirical research on business fixed investment has a long history, but the modern debate begins with the work of Aftalian (1909), Clark (1917) and Fisher (1930). Aftalian and Clark observed that business investment is highly correlated with changes

in business output – providing support for the early "accelerationist" school – while Fisher's neoclassical theory highlighted the importance of the trade-off firms make at the margin between the cost of raising more money and the benefit of the profit generated by an extra machine. The debate between these two schools provides a useful introduction to a review of the literature relating tax policy to investment. Many observers even recently [e.g., Clark (1993)] have argued that tax policy likely does not significantly affect investment, and the differing points of view inevitably harken back to the accelerationist debate.

3.1. *Neoclassical theory: a reprise*

The simplest neoclassical argument is that a firm weighs the costs and benefits of purchasing a machine today and holding it for one period. The firm invests when the benefits exceed the costs.

The economic logic underlying the user-cost concept we derived in Section 2 can be demonstrated with a simple example. Let the firm operate for one year, after which it will sell any capital it has acquired and close down. The firm will buy new capital at the beginning of period t at price q_t and sell it at the beginning of the next period at a price q_{t+1}. While the firm uses that capital, the machine depreciates from use. Again for simplicity, depreciation of capital takes place at the beginning of the period and the firm spends δq_t to replace the worn-out δ units of capital, and the increment to production, the marginal product of capital MPK, takes place at the beginning of period t, is stored costlessly during the period, and is sold at the beginning of period $t+1$ for the value of MPK (assuming a constant price of output, normalized to unity). Following the general model we outlined earlier, ρ is the required rate of return for investors. The present value of the net cash flow follows from just adding up the pieces:

$$\textit{Present value of net cash flow from the machine} = -p_t - \delta p_t + \frac{\text{MPK}_{t+1} + p_{t+1}}{1+\rho}.$$

With decreasing returns, the firm will continue to purchase machines until the last machine just pays for itself. Thus, with depreciation, for the marginal investment, this expression yields:

$$\text{MPK}_{t+1} = p_t[\rho + \delta + \rho\delta - (\Delta p_{t+1}/p_t)],$$

where $\Delta p_{t+1}/p_t$ denotes the percentage capital gain or loss on the asset due to a change in its market price.

Ignoring the small interaction term $\rho\delta$ [7], the firm's cost of capital in use has three components: the first is the combined real cost of debt and equity financing, ρp_t, which

[7] In most formulations the expression $\rho\delta$ is omitted because it is assumed to be small, and also because it vanishes in continuous time.

incorporates the required real rate of return of bondholders and shareholders, each on an after-personal-tax basis; the second is the economic rate of decay of the capital with an unchanging relative price of new capital, δp_t; and the third is an offset due to an instantaneous real capital gain on the capital, $(\Delta p/p)p_t$. If there are diminishing returns, then the marginal product of capital decreases as more capital is purchased, so the demand for capital is inversely related to the user cost. If the required rate of return is in part determined by the interest rate, then the demand for capital will go up when interest rates go down. This formula is easily modified to include taxes on profits, and subsidies to capital such as an investment tax credit. When this is done, it is easy to show that today's user cost of capital is just C_t:

$$C_t = q_t(\rho + \delta - \Delta p/p)\frac{1 - t_c z - \mathrm{ITC}}{1 - t_c}.$$

This is the familiar formula derived by Hall and Jorgenson (1967), which itself draws on the seminal work of Jorgenson (1963). Introduction of corporate taxes affects the user cost of capital in two ways, assuming the taxes are permanent. First, in the absence of tax deductions for depreciation and interest costs, an increase in the corporate income tax rate, t_c, increases the before-tax marginal product of capital necessary to yield an acceptable after-tax rate of return to investors, thereby increasing the user cost. Second, a higher corporate income tax rate increases the value of depreciation deductions and hence reduces the user cost. The multiplicative factor, $(1 - t_c z)/(1 - t_c)$, captures the combination of these two effects; on balance, the user cost is increased under current US tax law because expensing – or the immediate writeoff – of plant and equipment expenditures is not permitted (i.e., $z < 1$)[8]. If changes in tax are allowed, then these would enter the formula as well, in a straightforward extension that we leave as an exercise for the reader.

3.2. Early empirical results

Jorgenson (1963) investigated whether the neoclassical theory (building off of the user-cost formula above) could be used to describe aggregate fluctuations in business fixed investment in the United States. Moving from this equilibrium relationship to an empirical model, however, required a few more steps. Because output is determined by the choice of the capital stock, the theory does not relate the capital stock to a set of

[8] Third, a higher corporate tax rate increases the value of interest deductions and hence, all else being equal, reduces the real cost of debt financing. Given realistic parameter values, however, the first effect dominates: On balance, corporate taxes increase the user cost or the minimum pre-tax marginal product of capital necessary to yield an acceptable real rate of return to investors. As a consequence, corporate taxes in the United States diminish the incentive to invest. See Cohen, Hassett and Hubbard (1999) for a full discussion of this and related points.

exogenous variables[9]. Rather, it expresses a relationship among endogenous variables that holds in equilibrium. Indeed, given an assumption about the technology that turns capital into output, the theory does not define an investment relationship, that is, the *flow* of capital, but rather describes only the equilibrium *stock* of capital[10]. Jorgenson moved to an "investment" specification by defining a firm's "desired" capital stock, K^*, as output, divided by the user cost, Y/c, and then assuming that the firm gradually approached this desired stock over time. As opposed to relying on adjustment costs (cf. the analytical discussion in Section 2), Jorgenson assumed that the rate ω at which the firm closed the gap between its actual and desired stocks was given exogenously, and did not affect the level of the "desired" stock. These assumptions yielded the estimating equation

$$I_t = \sum_{i=0}^{T} \omega_i (K_{t-i} - K^*_{t-1-i}) + \delta K_{t-1}. \tag{15}$$

Hall and Jorgenson (1967) originally used such a model to explain aggregate investment, and concluded that it described the data well. Eisner and his collaborators later pointed out that the model they estimated – recognizing that K^* was the ratio of output to the user cost – could be capturing accelerator effects, which had long been known to be strong explanatory factors for investment. In particular, if one constrained the user cost to be a constant, one could rewrite Equation (15) as

$$I_t = \sum_{i=0}^{T} \omega_{i,\mathrm{uc}} (Y_{t-i} - Y_{t-1-i}) + \delta K_{t-1}, \tag{16}$$

which is a form of an accelerator model. When critics of Hall and Jorgenson isolated the separate contribution of the user cost to explaining investment, they found the user-cost effect to be negligible [see Eisner (1969, 1970), Eisner and Nadiri (1968) and Chirinko and Eisner (1983)].

Subsequently, while the neoclassical school may have had the theoretical high ground – because the user cost is clearly not constant over time – empirical implementations of neoclassical models using time-series data have not been successful. The time-series evidence has always revealed that lags of output are highly correlated

[9] To be more specific, Jorgenson assumed that the revenue function of the firm was Cobb–Douglas and that the firm set marginal revenue (product of capital) equal to the user cost in order to maximize profits.

[10] For example, Haavelmo (1960) writes "The demand for investment cannot simply be derived from the demand for capital ... I think the sooner this naive, and unfounded theory of the demand for investment schedule is abandoned, the sooner we shall have a chance of making some real progress in constructing more powerful theories to deal with the capricious short-run variations in the rate of private investment" [quoted in Jorgenson (1967), p. 133].

with investment, while interest rates and tax variables have generally provided very limited additional explanatory power. Models emphasizing the net return to investing are defeated in forecasting "horse races" by *ad hoc* models, and structural variables are frequently found to be economically or statistically insignificant[11].

As such negative evidence mounted, many economists became convinced that interest rates, taxes, and the other components of the user cost do not help predict investment behavior because firms do not pay attention to these variables. By contrast, corporate decision makers cite the user cost as an important concern when evaluating investment projects [see Hassett and Hubbard (1999)].

Hence, while by the late 1960s the neoclassical model developed by Jorgenson and others had become the standard model for studying the response of investment to tax policy, practical problems remained. On the one hand, the neoclassical approach offers a theoretical link between tax-policy parameters – the corporate tax rate, the present value of depreciation allowances, and the investment tax credit – and investment through the user cost of capital[12]. On the other hand, the empirical evidence suggested that the more rigorous theory did not improve the econometrician's ability to explain aggregate investment fluctuations or the response of business investment to changes in tax policy. Indeed, the tax-policy variables were often found to have no effect at all on investment.

3.3. Contemporary empirical tests of neoclassical models

These facts presented two challenges for empirical research linking tax policy and investment. First, a theory needed to be derived which described why yesterday's output appeared to be important empirically, even though any benefit of investment will occur in the future. Second, a coherent explanation of why investment by firms might actually respond to changes in interest rates and tax rates, while aggregate investment does not appear to in time-series data. This section summarizes efforts to address the first challenge; the following section turns to the second.

Motivated by the hope that the simplest neoclassical models failed to explain investment fluctuations because they were too stylized, substantial energy was devoted to the task of extending these models to incorporate more realistic assumptions in the 1970s and early 1980s[13]. By the late 1980s, substantial progress was made addressing

[11] See, e.g., Bosworth (1985), Bernanke, Bohn and Reiss (1988), and the survey in Chirinko (1993). The often poor empirical performance of Q models has led some researchers to abandon the assumptions of reversible investment and convex costs used in testing neoclassical models in favor of approaches based on lumpy and "irreversible" investment. See, e.g., the discussions and reviews of studies in Pindyck (1991), Dixit and Pindyck (1994) and Hubbard (1994).

[12] Feldstein (1982), for example, explored the effects of effective tax rates on investment in reduced-form models; for a critique of this approach, see Chirinko (1987). We return to this debate below.

[13] Eisner and Strotz (1963) offer an early discussion of adjustment costs. The theory was developed and extended by Lucas (1967, 1976), Gould (1968), Treadway (1970), Uzawa (1969), Abel (1980) and

the first of the two challenges. The most significant step occurred when theorists explicitly incorporated costs of adjusting the capital stock into their models. According to these new theories, firms face very large costs if they attempt to make very large instantaneous changes in their production technologies, and such costs fall significantly if the firm changes its capital stock gradually. This new assumption provided a link between what the firm was doing yesterday and what it plans to do tomorrow that was absent in the first neoclassical models. According to these models, investment is forward-looking, and based upon rational expectations of future variables that affect profit at the margin, but it also depends on how much capital is already on hand. Because firms base their expectations of future variables in part on their observations of the past, researchers identified a link between a set of lagged variables and current investment. Anything that helps predict future market conditions might plausibly matter in investment regressions. Hence a correlation between past output growth and future "fundamentals" could be used to rationalize a strong correlation between current investment and past values of the growth of output.

Such new investment models emphasizing the net return to investment, but with adjustment costs, have yielded complementary empirical representations. Each begins with the firm maximizing its net present value. The first-order conditions with respect to investment and capital lead to an Euler equation describing the period-to-period optimal path of investment. Investment today depends on prices, taxes, interest rates, and what you expect investment tomorrow to be. Abel and Blanchard (1986) solved the Euler equation and developed an estimating equation that relates investment to its expected current and future marginal revenue products of capital. Alternatively, effects of tax parameters may be estimated from the Euler equation itself [see, e.g., Abel (1980) and Hubbard and Kashyap (1992)]. As in Auerbach (1989a) and Abel (1990), investment can be expressed in terms of current and future values of the user cost of capital and, under some conditions, expressed in terms of average q, which is the market value of the firm divided by the replacement cost of capital. This approach was suggested initially by Tobin (1969), with the necessary conditions supplied by Hayashi (1982)[14].

The equilibrium *marginal q* is related to the price of investment goods, tax parameters, and adjustment costs. If we assume that the adjustment function is quadratic,

$$C(I_{i,t}, K_{i,t-1}) = \frac{\omega}{2}\left(\frac{I_{i,t}}{K_{i,t-1}} - \mu_i\right)^2 K_{i,t-1},$$

Hayashi (1982). Researchers have generally assumed convex costs of adjusting the capital stock; the idea is that it is more costly to implement a given increment to the capital stock quickly rather than gradually. We discuss alternative assumptions about adjustment costs in Section 4.

[14] Hayashi (1982) provided the conditions required to equate marginal q with average Q, which is observable because it depends on the market valuation of the firm's assets. Summers (1981) incorporated additional tax parameters in the Q model.

where μ is the steady-state rate of investment and ω is the adjustment-cost parameter, then Equation (8b) (abstracting from personal taxes) can be rewritten as an investment equation:

$$\frac{I_{i,t}}{K_{i,t-1}} = \mu_i + \frac{1}{\omega}\left[\frac{q_{i,t} - p_t(1-\Gamma_{i,t})}{(1-\tau_t)}\right]. \tag{17}$$

Equation (17) offers a convenient way of estimating the responsiveness of investment to neoclassical variables, including tax parameters, but there is a complication: marginal q is unobservable. Following Hayashi (1982), if the firm is a price taker in input and output markets and the production function exhibits constant returns to scale, marginal q equals average q, defined for each firm as tax-adjusted q (denoted below by Q):

$$Q_{i,t} = \frac{V_{i,t} + B^{c}_{i,t} - A_{i,t}}{K^{R}_{i,t}},$$

where V is the market value of the firms' equity, B^{c} is the market value of the firm's debt, A is the present value of depreciation allowance on investment made before period t, and K^{R} is the replacement value of the firm's capital stock (including inventories).

When used to explain the time-series movements of investment, however, Q models proved very disappointing as well. The basic accelerator model, that depends only on output, did just as well as, if not better than, the Q theory in forecasting horse races. Moreover, parameter estimates for the new models tended to be wildly implausible. The estimated coefficient on Q, γ indicates the speed with which firms can adjust their investment to its target or optimal level. If the estimated Q coefficient is very small, then investment does not respond quickly to Q values different from the long-run equilibrium value. The very small Q coefficient reported in the literature often implied that the costs of adjustment incurred when installing a new machine were larger than the purchase price of the machine itself.

Researchers usually estimated such models using either ordinary least-squares or generalized-method-of-moments techniques with instrumental variables. Cummins, Hassett and Hubbard (1994, 1996) note that conventional estimated values of γ in firm-level panel data for the United States or for other countries are very small, ranging from 0.01 to 0.05, implying marginal costs of adjustment of between one and five dollars per dollar of investment. Such estimates, which have emerged in many empirical studies [see, e.g., Summers (1981), Salinger and Summers (1983) and Fazzari, Hubbard and Petersen (1988a)], imply very small effects of permanent investment incentives on investment. Applications of the alternative approaches to time-series data, while promising, continued to suffer by comparison to accelerator models.

This work completed the first wave of responses to the neoclassical failure. The second wave of responses explored the alternative specifications using much

richer data sources than had generally been used in the past. Before discussing these approaches and evaluating their contributions, we begin by presenting several charts that illustrate the empirical difficulties confronted in estimating time-series relationships between tax variables and investment.

3.4. *Lessons from the time-series data*

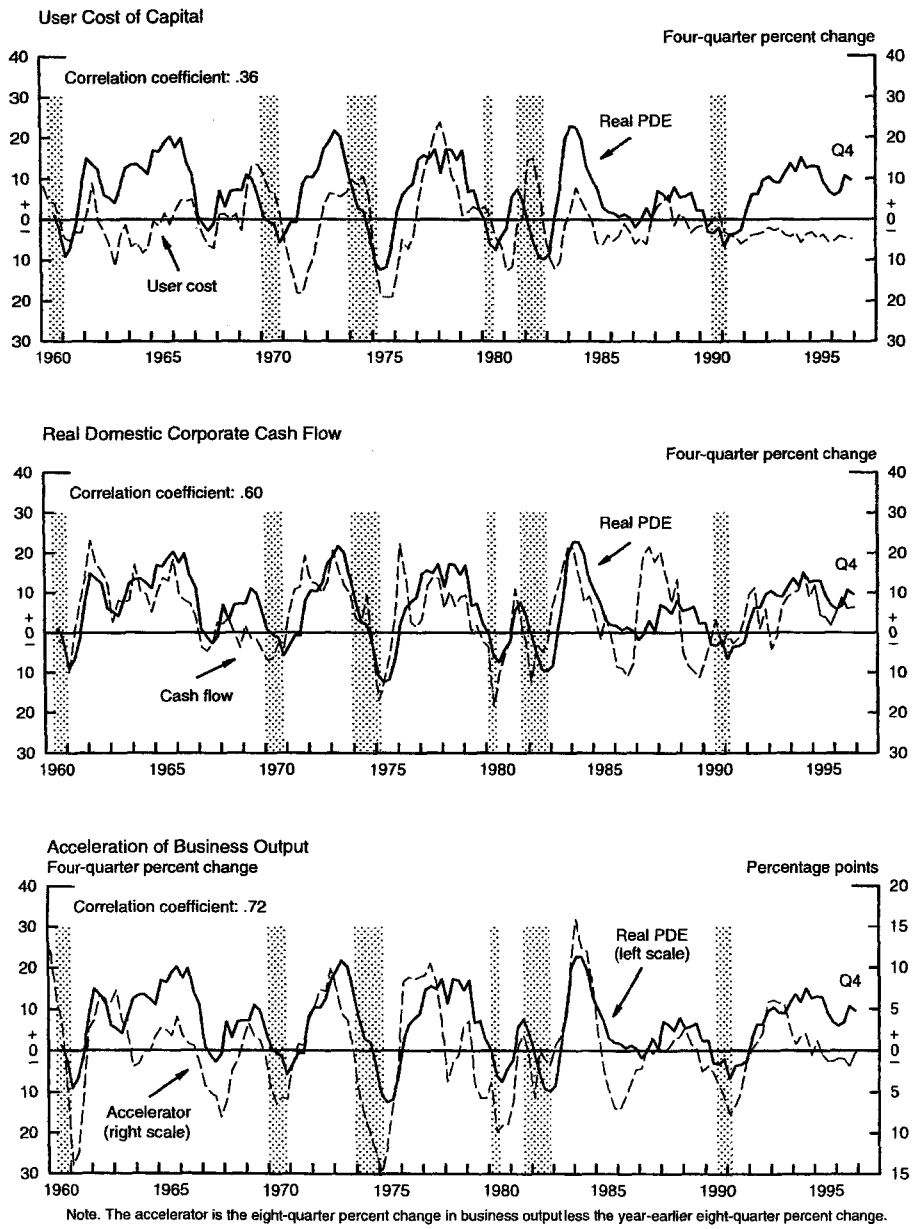
Figure 4 plots aggregate US equipment investment against several investment "fundamentals". The top panel shows the comovement of investment and the user cost of capital, which is based on the corporate AAA bond rate and historical tax laws. The series rarely move together in an obvious way, and the correlation since 1960 is a statistically insignificant 0.36. The second panel illustrates the strong comovements between investment and corporate cash flow, which here is measured as corporate profits plus interest payments and depreciation. The two series are roughly coincident, and the correlation over time is a highly significant 0.60. The bottom panel illustrates the "accelerator" effect, which relates changes in the growth rates of output and equipment spending. As with cash flow, the correlation is large, 0.72, and highly significant, and the coincidence of the series is visually striking.

While one should be cautious interpreting such correlations formally, they nonetheless suggest clear patterns. Aggregate equipment investment varies significantly over the business cycle, and neither lags or leads the cycle; it is highly correlated with other variables that are also highly procyclical. The time-series correlation between investment and the user cost, on the other hand, is quite weak. Figure 4 can be thought of as a visual summary of the early investment literature: Accelerator effects are strong and obvious; user-cost effects appear weaker and more subtle.

For purposes of illustration we focus on equipment investment, in large part because empirical attempts to model investment in structures have been more disappointing. Figure 5, which repeats Figure 4 with the relevant "fundamentals" related to the growth rate of investment in nonresidential structures, illustrates the problem. Structures investment is less clearly correlated with all of the "fundamentals". The correlation with the user cost is insignificant and has the incorrect sign, the correlation with cash flow is about one-fourth of that between cash flow and equipment investment, and the accelerator effect, while still noticeable, is significantly weaker.

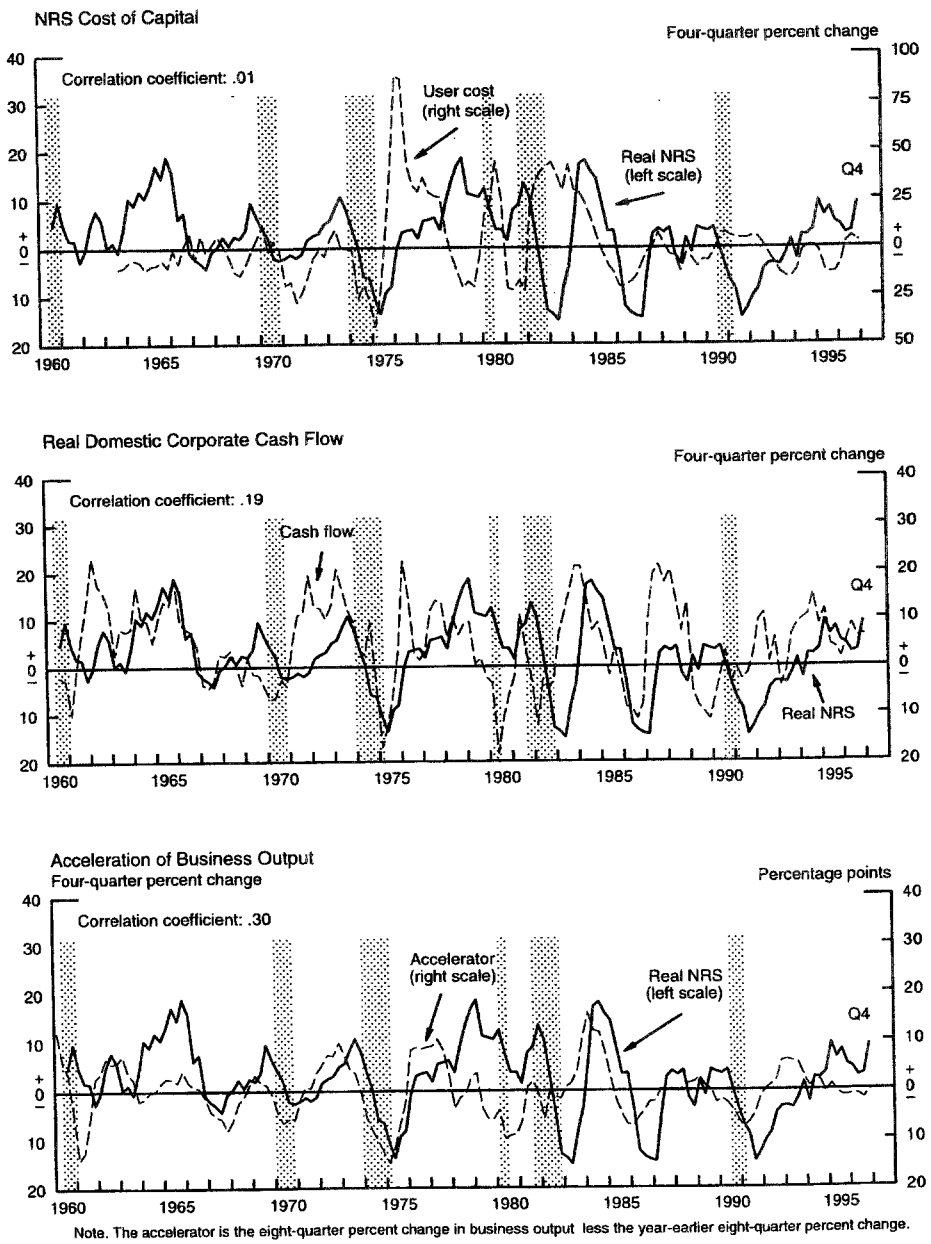
Figure 6 depicts the correlation of aggregate business fixed investment with Q [15]. The top panel compares the level of real investment to the level of Q. Clearly, the low-frequency movements in the two series are not highly correlated. The bottom panel relates the growth rates of these two series. Here it appears that growth in Q

[15] The measure of Q plotted here is constructed from data from the Federal Reserve's Flow of Funds Accounts.



Note. The accelerator is the eight-quarter percent change in business output less the year-earlier eight-quarter percent change.

Fig. 4. Fundamental determinants of equipment spending.



Note. The accelerator is the eight-quarter percent change in business output less the year-earlier eight-quarter percent change.

Fig. 5. Fundamental determinants of NRS spending.

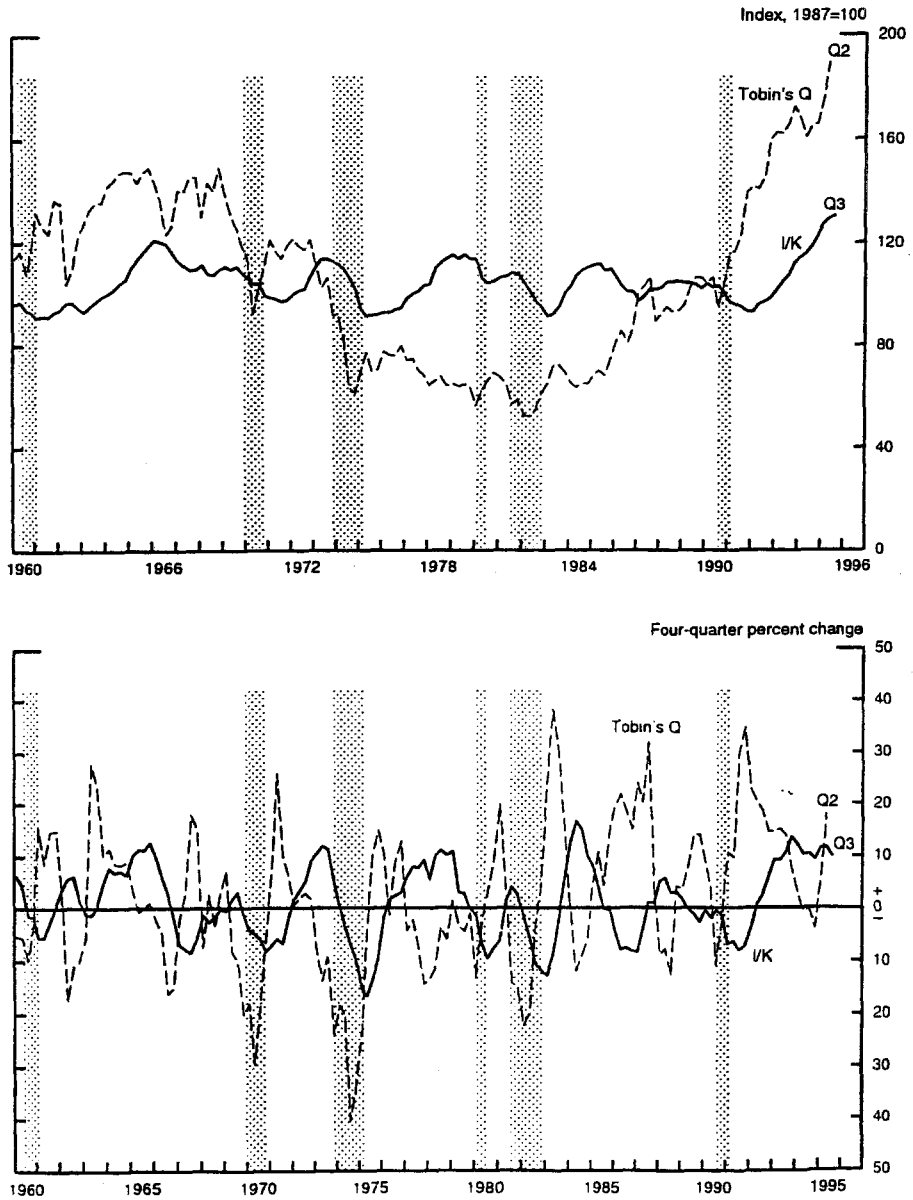


Fig. 6. Tobin's Q and the I/K ratio.

leads growth in investment somewhat, although the relationship is weak, and the contemporaneous correlation is actually negative[16].

Does this mean that tax variables do not affect investment as predicted by the neoclassical model? Movements of the aggregate variables – including investment – are determined simultaneously, of course, and disentangling the marginal impact of a single driving variable is difficult (if not impossible) using time-series data. For example, suppose that aggregate demand increases exogenously for some reason. This shift might lead firms to be more optimistic about their sales prospects and to purchase more investment goods; it might also be expected – at least in the short run – to lead to higher interest rates. If you then examine the correlation between investment and the interest rate, you might even find that the sign is the opposite of that predicted by the theory. While an instrumental-variables procedure might allow us to overcome this simultaneity problem, the estimator is only as good as the instruments, and it is difficult to imagine an appropriate set of instruments for this application.

To summarize, the tendency for a number of aggregate variables to move together over the business cycle makes it difficult to isolate effects of individual fundamentals on investment using time-series data. Even if investment is very responsive to tax policy, it might appear not to be in the aggregate data, since so many other important determinants of investment are moving over the business cycle as well. Hence a partial-equilibrium investment demand approach might have very little power to explain aggregate investment fluctuations or links between tax policy and investment. Microeconomic data, however, provide a rich additional source of variation, and it is to the microdata studies that we now turn.

4. New identification strategies in empirical research

As we discussed earlier, one reason the data do not appear to favor neoclassical models is a simultaneous-equations problem: If, on the one hand, the data were dominated by exogenous increases or decreases in the real interest rate, then the user-cost movements would lead investment to decrease or increase, respectively. If, on the other hand, investment rises with positive "animal spirits", then higher investment demand puts upward pressure on the real interest rate. Hence, to the extent that data incorporate both exogenous changes in the real interest rate and in the intercept of the investment function, the positive relationship between investment and the user cost of capital because of shifts in the investment function may dominate the

[16] If the growth rate of business fixed investment is regressed on many lags of the growth rate of Q, the sum of the coefficients is about 0.1, implying that a 20-percent increase in the growth rate of Q would lead to a prediction of about a 2-percent higher growth rate of business fixed investment. Cochrane (1991) finds significantly larger effects of the growth of Q on the growth of total private investment. The results differ because Cochrane's measure of investment includes residential investment, which he finds is more highly correlated with stock market fluctuations.

hypothesized negative relationship between investment and the user cost of capital. In this case, the estimated coefficient on the user cost of capital will be "too small", leading to estimated adjustment costs that are "too large". Such simultaneity increases apparent accelerator effects, because positive shifts of the investment function raise both investment and output. Controlling for these effects has been the goal of the second wave of research.

4.1. Using cross-sectional variation to identify tax effects

In principle, this simultaneity problem in the estimation of neoclassical models can be tackled by the use of instrumental variables. Conventional instrumental variables (including lagged endogenous variables or sales-to-capital ratios) have not proven very helpful. Major tax reforms, however, arguably offer periods in which there is exogenous *cross-sectional* variation in the user cost of capital or tax-adjusted q. Auerbach and Hassett (1991) and Cummins, Hassett and Hubbard (1994, 1995a, 1996) demonstrate that major tax reforms are also associated with significant firm- and asset-level variation in key tax parameters (such as the effective rate of investment tax credit and the present value of depreciation allowances). Hence tax variables are likely to be a good instrument for the user cost or Q during tax reforms. Using a related approach, Carroll, Holtz-Eakin, Rider and Rosen (2000) show that cross-sectional variation in change in personal tax rates following the Tax Reform Act of 1986 is associated with variation in investment by small businesses.

The variation across assets is large within most years, as is the time-series variation. In addition, the relative treatment of different assets changes somewhat over time. For example, following the removal of the investment tax credit and the reduction of the corporate tax rate by the Tax Reform Act of 1986, the cross-sectional variation in the tax treatment of capital assets fell, consistent with the Act's stated goal to "level the playing field".

Auerbach and Hassett (1991) and Cummins, Hassett and Hubbard (1995a) use vector autoregressions to forecast investment in the year following a tax reform, and then compare the forecast errors for each of the assets to the changes in the user cost for that asset. Auerbach and Hassett (1991) provide a proof that a regression of one error against the other can capture the underlying structural parameters of the investment model if the firm knows that a tax reform is coming, but the econometrician can only predict the tax reform with error. They argued that these assumptions were reasonable for the 1986 act, as it included phased-in changes to tax rates that were known by firms with certainty. Both sets of authors find a clear negative correlation in these surprises.

Table 3 shows the significance of using exogenous tax changes to identify changes in Q [17]. Taken from Cummins, Hassett and Hubbard (1996), it presents estimates of

[17] Cummins, Hassett and Hubbard (1994, 1995a) also use this approach in a user cost model. For US data, they estimate a user-cost coefficient of about –0.65.

Table 3
Estimates of tax-adjusted Q model for fourteen countries[a]

Country	Conventional panel data estimated coefficient on Q[b]	Estimated coefficient with contemporaneous tax instruments[c]
Dependent Variable: I/K		
Australia	0.050	0.289
	(0.019)	(0.153)
Belgium	0.103	0.587
	(0.044)	(0.422)
Canada	0.041	0.521
	(0.009)	(0.127)
Denmark	0.104	0.765
	(0.085)	(0.308)
France	0.085	0.388
	(0.042)	(0.116)
Germany	0.095	0.784
	(0.040)	(0.296)
Italy	0.051	0.180
	(0.018)	(0.120)
Japan	0.029	0.086
	(0.008)	(0.035)
The Netherlands	0.069	0.633
	(0.044)	(0.150)
Norway	0.069	0.512
	(0.031)	(0.295)
Spain	0.044	0.404
	(0.028)	(0.233)
Sweden	0.051	0.293
	(0.047)	(0.169)
United Kingdom	0.062	0.589
	(0.013)	(0.078)
United States	0.048	0.650
	(0.006)	(0.077)

[a] Source: Calculations in Cummins, Hassett and Hubbard (1996) using Global Vantage data; standard errors are in parentheses.

[b] See Table 5 in Cummins, Hassett and Hubbard (1996), GMM estimates. Instruments include twice- and thrice-lagged values of Q, (I/K), and the ratios of cash flow to capital.

[c] See Table 7 in Cummins, Hassett and Hubbard (1996), GMM estimates. Instruments include twice- and thrice-lagged values of (I/K) and the ratio of cash flow to capital, twice-lagged value nontax components of q, and contemporaneous values of tax parameters.

a simple equation relating the ratio of investment to capital to Tobin's Q during major tax reforms in 14 countries over the 1980s; firm-level data are taken from Compustat's Global Vantage. Using contemporaneous tax variables as instrument during major tax reforms, Cummins, Hassett, and Hubbard estimate the coefficient on Q to be 0.65 for the United States, compared with a paltry 0.048 under conventional estimates. They obtained similar estimates for each of the other major US tax reforms in the postwar period using data from Compustat [Cummins, Hassett and Hubbard (1994)]; focusing on the Tax Reform Act of 1986, Auerbach and Hassett (1991) found similar coefficients using asset-level data and cross-sectional variation in the user cost. Chirinko, Fazzari and Meyer (1999) find significant effects, but also show that some specifications suggest smaller elasticities.

Since the identification of these large effects depends so crucially on cross-section variation, it may be that asset substitution is important, but that the level of aggregate investment responds little to tax policy. We return to this question below.

Subsequently, empirical researchers have offered three general explanations of the failure to estimate significant tax effects on investment – (1) measurement error in fundamental variables, (2) misspecification of costs of adjusting the capital stock, and (3) the importance of capital-stock heterogeneity. All three research programs have contributed to our understanding of the responsiveness of investment to changes in the net return to investing and have reached similar conclusions about the likely effects of tax policy for some important cases.

4.2. Measurement error in fundamental variables

These alternative estimation approaches argued that the presence of measurement error strongly affects results based on time-series variation. An important recent note [Goolsbee (2000a)] has documented that such measurement error is indeed important. A number of alternative techniques have been suggested to address this measurement error, including: (1) statistical corrections, (2) avoiding the use of Q or user-cost representations, (3) using new proxies for Q, (4) focusing on periods or frequencies in which firm variation in fundamental variables is less subject to measurement error, and (5) modifying assumptions about the financial frictions firms face. Each possibility is considered in turn below.

There are at least two problems in measuring Q that might affect estimated adjustment costs. First, to the extent that the stock market is excessively volatile, Q might not reflect market fundamentals. Second, the replacement value of the capital stock in the denominator of Q is likely to be measured with error. Griliches and Hausman (1986) argue that measurement error will lead to different biases among potential estimators that are similar in that they control for firm-specific effects, but differ in their signal-to-noise ratios, making it possible to place bounds on the importance of measurement error. Following a suggestion by Griliches and Hausman (1986), Cummins, Hassett and Hubbard (1994) estimate a Q model using first differences and longer differences (as opposed to the conventional fixed-effects,

within-group estimator) to address measurement-error problems. Their estimated adjustment costs decline significantly in the long differences confirming that standard specifications may have important measurement-error problems.

In a time-series setting, Caballero (1994) pursues an alternative estimation strategy, based on a suggestion by Stock and Watson (1993). Stock and Watson show that, in a time-series setting, small-sample bias is reduced if leads and lags of integrated regressors are included in a regression model. Caballero argues that small-sample biases of typically employed time-series estimation procedures are particularly severe when estimating adjustment-cost models, and he shows that elasticities will generally be biased downward. This is because a "frictionless" capital stock such as Jorgenson's optimal K would fluctuate more than the observed capital stock if there are adjustment costs. (If there are no adjustment costs, then capital would fluctuate with the frictionless stock.) A regression model that uses a user cost to measure the "target" or frictionless capital stock will therefore have an error term that is negatively correlated with the target capital stock, and depends on adjustment costs. Caballero demonstrates that the Stock and Watson method solves this problem, and uses it to estimate a long-run elasticity of investment with respect to the user cost of approximately unity. This is much larger than the early estimates, but roughly consistent with the other studies summarized in this section.

Another approach departs from the strategy of using proxies for marginal q, and relies on the firm's Euler equation to model the investment decision. (As long as one makes the same assumption about technology and adjustment costs, the Euler equation can be derived from the same model as the conventional Q or user cost of capital models). By not relying on the "investment function" representation, one can sidestep problems of measuring marginal q.

Tests following this approach have frequently used panel data on manufacturing firms to estimate the Euler equation [Pindyck and Rotemberg (1983), Shapiro (1986), Gilchrist (1991), Whited (1992), Bond and Meghir (1994), Hubbard, Kashyap and Whited (1995)]. Studies using Compustat data for the United States are unable to reject the frictionless neoclassical model for most firms, and the estimated adjustment-cost parameters are more reasonable than those found in estimates of Q models. For example, Hubbard, Kashyap and Whited (1995) report Q-coefficient estimates between 1 and 2.2. Very similar estimates are reported for European manufacturing firms by Cummins, Harris and Hassett (1995b) and for investment in overseas subsidiaries of US multinational corporations in Cummins and Hubbard (1995).

The measure of average Q used as a proxy for marginal q in most empirical studies is constructed as the ratio of the market value of the financial claims on the firm (equity and debt) to the replacement cost of the firm's capital stock. The third approach bypasses using financial variables as proxies for marginal q by forecasting the expected present value of the current and future profits generated by an incremental unit of capital – that is, the expected value of marginal q – an idea developed by Abel and Blanchard (1986) and applied to panel data by Gilchrist and Himmelberg (1995, 1998). Using such an approach, Gilchrist and Himmelberg (1995, 1998) report estimates

of adjustment costs that are roughly consistent with the Euler-equation estimates discussed above.

To summarize, a variety of empirical implementations of the neoclassical model with convex adjustment costs have attempted to mitigate measurement error and other econometric problems in conventional OLS and GMM estimates using panel data. The methods described above generally yield estimates that imply marginal costs of adjustment in the range of \$0.10 per dollar of additional investment (using the estimate in Cummins, Hassett and Hubbard (1995a) as a benchmark), and elasticities of investment with respect to the user cost of capital between −0.5 and −1.0.

4.3. An alternative interpretation: misspecification of adjustment costs

The empirical studies just mentioned accept the conventional belief that costs of adjusting the capital stock are convex. The *Q,* user cost of capital, and Euler-equation approaches can all be derived from the same intertemporal maximization problem, given common assumptions about technology, competition, and adjustment costs. An important recent line of inquiry focuses on modeling and testing the effects of irreversibility and uncertainty on firms' investment decisions [see, e.g., the excellent survey by Dixit and Pindyck (1994)][18]. If these effects are important, then there may be ranges of values for fundamentals in which tax policy has little or no effect on investment, and knowledge of which range firms are operating in is a prerequisite for policy analysis. Finally, these models can possibly explain the puzzle of why firms report in surveys that they use such high hurdle rates [see Summers (1987)].

Neoclassical models implicitly assume that there is an efficient secondary market for capital; hence irreversibility poses no problem. If a firm purchases a machine today, and the output market turns sour in the future, the firm can recoup the purchase price of the machine at that time. Returning to the discussion in Section 3.7, if investment is irreversible, then the firm faces the chance that it cannot sell the machine in the future. In this setup, there is a gain to delaying investment and allowing the random price process to move either into a region far enough above the neoclassical "break-even" point that the probability of the "bad state" becomes low enough, or into a region where it clearly does not make sense to purchase the machine. An investment extinguishes the value of the call option of delay, an option that has positive value when prices are uncertain. In this approach, the value of the lost option is a component of the opportunity cost of investment. In the terminology of the Q framework, the threshold criterion for investment requires that marginal q exceed unity by the value

[18] The seeds of this literature are much older. For example, Rothschild (1971) writes: "Convex cost-of-adjustment functions may help to explain why Rome was not built in a day. However, there is no clear saving and may be some loss to spreading the work of installing a button on a shirt over several weeks". While not a model of irreversible investment, his "bang-bang" model of investment provides an early example of a model with an alternative adjustment cost function.

of maintaining the call option to invest. As a consequence, high "hurdle rates" may be required by corporate managers making investment decisions.

Indeed, at least part of the interest in option-based investment models has been the problem raised in many time-series studies that indicated that the response of investment to changes in Q or the user cost of capital are implausibly small, implying, perhaps, that there are regions wherein Q varies but investment does not. In addition, it is not difficult to suggest examples of nonconvex adjustment costs – such as retooling in automobile plants or the adoption of more energy-efficient kilns in cement plants.

Abel and Eberly (1994, 1996a,b, 1999) provide a general framework that encompasses irreversibility, fixed costs, and a wide array of alternative adjustment-cost specifications. They show that, under certain conditions, the investment behavior of firms can be characterized by three distinct regimes: (1) regime in which gross investment is positive; (2) regime in which gross investment is zero; (3) regime in which gross investment is negative. This contrasts with the linear relationship predicted by the quadratic cost model. The responsiveness of investment to fundamentals differs across regimes, and their more general model predicts that there is a region in which gross investment will stay zero for a range of unfavorable values of Q. Because this model nests the more traditional q models, it provides a useful empirical framework.

Researchers are beginning to study the impact of alternative adjustment-cost assumptions within structural investment models with panel data. Barnett and Sakellaris (1998) use Compustat data to investigate the implication of the model of Abel and Eberly (1996b) that investment alternates between regimes of insensitivity to Q and regimes of responsiveness to Q. The region of inactivity should be close to the region for which the model predicts that investment is negative. Because the thresholds for these regions are unknown, conventional asymptotic distributions do not apply. Barnett and Sakellaris use a statistical framework that allows them to estimate the threshold points and the coefficients on Q simultaneously in the different regions given the threshold points. They find evidence of a nonlinear relationship between investment and Q; in particular, they estimate the largest responsiveness of investment to Q for *low* values of Q, and the smallest for very *high* values of Q. On average, they estimate that the elasticity of investment with respect to Q is about unity, but that the aggregate elasticity varies considerably over time, depending on the average level of Q. Barnett and Sakellaris argue that their results imply that adjustment cost may not be quadratic, but that the most likely cause is not firms' inability to disinvest, but rather, their reluctance to make "large" changes.

Barnett and Sakellaris's results are not necessarily inconsistent with the story of measurement-error. Some firms in their Compustat universe have values of average Q that are astronomical, presumably because the capital-stock measure is missing important goodwill or human-capital components. If one accepts that Q is a poor measure of fundamentals for these firms, then the result that investment does not respond as much to Q for these firms is not surprising. In the more "normal" range of Q values, the investment response seems to accord well with the predictions of the convex adjustment cost model.

Caballero, Engel and Haltiwanger (1995) explore adjustment costs in a more general framework. Using a subset of 7000 US manufacturing plants from the Census Bureau's Longitudinal Research Database (LRD), they explore whether cross-sectional patterns of investment are consistent with symmetric, convex adjustment cost models, or whether the data imply that there are nonconvexities[19]. They proceed in two steps. First, they assume that there are no adjustment costs and that the Jorgensonian model adequately describes a firm's "desired" capital stock (K^*)[20]. They then compare in each period a firm's beginning-of-period capital stock to its desired stock and call the difference ($K^* - K_{t-1}$) "mandated investment." Second, they explore how firms actually adjust their capital stocks. In this step, they find that the relationship between actual and mandated investment is highly nonlinear. If mandated investment is negative, then firms do not quickly adjust their capital stocks downward. If mandated investment is small and positive, then firms also do not respond very much. If mandated investment is very large, then firms adjust their capital stocks very quickly. They conclude that a kinked adjustment or (S,s) model, in which firms have a range of inaction, and only adjust their capital stocks to their desired levels when the gap between current and desired capital stock is "large enough" offers a good description of the data[21].

Using firm-level data from Compustat, Abel and Eberly (1996b) estimate that the relationship between investment and fundamental determinants (Q and the tax-adjusted price of capital goods) is concave; that is, the response of investment to fundamental determinants is positive, but monotonically declining. The results of Abel

[19] In earlier work, Doms and Dunne (1994) report that plant-level investment data exhibit skewness and kurtosis that is consistent with investment being somewhat "lumpy".

[20] To calculate the desired capital stock for each firm, Caballero, Engel and Haltiwanger use plant-level output data, and two-digit Jorgensonian user costs constructed from the tax data used in Cummins, Hassett and Hubbard (1994) and Goolsbee (1998).

[21] Doyle and Whited (1998) provide important additional evidence suggesting that the (S,s) model may provide an important microfoundation for future macro-investment work. The authors explore the relationship between deviations of optimal from actual capital across industries and the proportion of industry risk that is idiosyncratic, a relationship first explored by Bertola and Caballero (1990). They show that in (S,s) models this ratio is negatively correlated with persistence in industry-aggregate deviations from optimal capital. That is, if most shocks are idiosyncratic in a given industry, then the shocks will cancel out at the industry level, and the "frictionless" model may be a reasonable description of aggregate fluctuations in that industry. If most shocks affect the industry as a whole, then the industry might look like an individual (S,s) firm, with highly persistent deviations of actual from desired capital. Doyle and Whited construct a measure of risk and industry measures of capital and show that the greater idiosyncratic the risk in an industry, the more fleeting are the deviations from optimal aggregated capital. This evidence suggests that traditional models may describe incompletely short-run dynamics surrounding tax reforms.

Goolsbee and Gross (1997) analyze aircraft replacement by airlines, and find clear evidence of nonconvexities, with firms demonstrating an area of inaction, but with quadratic adjustment costs conditional on making an investment. They show that aggregation obscures the nonconvexities, and biases upward estimates of adjustment costs.

and Eberly suggest that the distribution of tax-adjusted Q or the user cost of capital may be a determinant of aggregate investment. However, the caution that applied to Barnett and Sakellaris conclusions applies here as well: Large observed values of Q may not coincide with high levels of investment because the high Q values reflect mismeasurement, rather than extraordinary fundamentals.

Caballero, Engel and Haltiwanger also illustrate how to construct aggregate implications from their microeconomic results. Integrating over the microeconomic distribution of plants, they calculate a predicted aggregate elasticity of investment with respect to the user cost of capital. The estimates of this elasticity vary considerably over time: If, on the one hand, many plants are near the region for which mandated investment is very large, then small changes in the user cost can have large effects on aggregate investment. If, on the other hand, the bulk of the distribution of mandated investment is in the region of low responsiveness of investment to fundamentals, then changes in the user cost will have little impact. Like Cummins, Hassett and Hubbard (1994), they conclude that the aggregate elasticity of investment with respect to the user cost is between -0.5 and -1, and that tax reforms appear to have generally had large effects on investment. They caution, however, that the reforms have had large effects because they coincidentally occurred during periods in which the plant-level distribution of mandated investment was aligned in such a way to allow a large effect of changes in tax parameters. This would happen if, for example, investment tax credits were removed in booms, when mandated investment is very large, and an increase in the user cost can cause firms to cancel significant investment plans. As a consequence, Caballero, Engel and Haltiwanger argue that researchers must consult the microeconomic distribution of mandated investment before predicting the likely impact of future tax reforms on business investment[22].

Alternatively, Cummins, Hassett and Hubbard (1994, 1996) argue that recovering "reasonable" estimates of the response of investment to Q or the user cost of capital is easiest during periods in which large exogenous changes in the distribution of structural determinants occur, as during tax reforms. In response to the alternative interpretation that firms respond only to changes in fundamentals when these changes are large, Cummins, Hassett and Hubbard use firm-level data to investigate whether there was evidence of "bunching" of investment around tax reforms. They estimate transition probabilities among various ranges of (I/K) over the year prior to, the year of, and the year after the tax reform, and find no evidence that firms with large investment were likely to have lower investment in prior or subsequent years. Indeed, only a very tiny fraction of the sample was ever found to transit from high-investment to low-investment states.

[22] Because their mandated investment measure is the same as that in a frictionless neoclassical model, their tests – while suggestive – neither confirm or reject the presence of convex adjustment costs. First, mandated investment itself depends on adjustment costs. Second, if adjustment costs were present, mandated investment also depends on future values of tax parameters.

In part, the conclusions of these studies may differ because of differences in the level of aggregation. At a sufficiently fine level of disaggregation, all investment looks lumpy. The plant-level evidence suggests that investment appears lumpy, but the firm-level evidence does not corroborate this. However, there may be interesting differences between the investment behavior of plants and firms, as might be the case if, for example, managerial attention is limited and only a fraction of a firm's plants adjust their capital in a given year. Clearly, reconciling the plant-level and firm-level results is an important topic for future research.

4.4. The importance of heterogeneity

An alternative promising path has recently been opened up by research on capital heterogeneity [see Cummins and Dey (1998) and Goolsbee (2000b)]. While capital heterogeneity has largely been ignored in studies of investment behavior, one can, in principle, estimate a dynamic structural model in which different types of capital are interrelated in both the production and adjustment-cost technologies. The idea is that if one is going to shut down the plant to install new machines, then one might as well perform structural alterations at that time as well. This property produces a bunching of investment similar to that suggested by models of irreversibility; that is, estimates of adjustment costs are biased upward, and estimates of factor substitution in production are biased downward.

Goolsbee (2000b) argues that tax subsidies will change the relative price of "high-quality" capital, if "high quality" is interpreted to mean that the machine requires less future maintenance. Changes in the quality of machines could, in principle, alter significantly our perceptions concerning the effects of tax reform. In particular, Goolsbee shows that tax reforms appear empirically to be associated with large changes in the quality of machines purchased. To the extent that quality adjustments are not accounted for adequately by deflators, our perceptions about the effects of policy may be inaccurate.

4.5. Summary

Recent empirical studies appear to have reached a consensus that the elasticity of investment with respect to the tax-adjusted user cost of capital is between -0.5 and -1.0. Recent studies using convex costs of adjustment and studies using nonconvex costs of adjustments agree that the long-run elasticity of investment to the user cost is high by the standards of the early empirical literature. This range of estimated responses of investment to tax parameters is well above the consensus of only a few years ago, and suggests that investment tax policy can have a significant impact on the path of aggregate capital formation. One should be cautious, however, in moving from the microeconomic evidence to aggregate predictions. While Caballero, Engel and Haltiwanger (1995) demonstrate a technique for aggregating the microeconomic

distribution of firms to calculate aggregate investment demand, little continues to be known about the general-equilibrium effects of major policy changes.

5. Arguments for and against investment incentives

Consistent with neoclassical models of business fixed investment, research demonstrates both that tax incentives for investment are important components of the net return to investing, and that the long-term responses of investment to permanent tax incentives are large. A deeper policy question remains, however, of whether permanent incentives for investment are socially desirable even if such incentives increase the stock of business fixed capital. (We then address the question of the desirability of short-run incentives). Put differently, under what circumstances might one advocate distortionary investment incentives?

5.1. Tax reform could remove a distortion

Taxes increase the user cost of capital for both equipment and structures in most countries. The extension of expensing of capital investments would remove distortions associated with the current tax system. Indeed, the removal of capital-allocation distortions is one source of efficiency gains from proposals for broad-based consumption tax reform. An investment incentive could be designed to produce the same effect, although the tendency for these to be "targeted" in practice suggests that a uniform reduction in capital-income taxes might be difficult to obtain with this particular tool.

An additional argument for subsidies to equipment investment has been advanced by Judd (1997), who concludes that the optimal tax on equipment investment is negative. In Judd's model, capital-goods-producing industries are imperfectly competitive, and equipment prices contain significant markups[23]. These markups are analogous to tax distortions, and tax credits can return firms' input mix to the optimal social level if the government designs investment subsidies that equate the prices paid for different types of equipment to marginal cost. Judd (2001) reviews the impact of fundamental tax reform in a such model.

5.2. Investment incentives cause interasset distortions

In practice, investment tax credits have generally been applicable only to investments in equipment. One argument against such credits is that they introduce interasset distortions. If these interasset distortions are sufficiently large, gains from removing

[23] Goolsbee (1998) provides support for this view, while Hassett and Hubbard (1998) provide contradictory evidence.

the intertemporal distortion from the higher capital-income tax on investment might be eliminated.

Auerbach (1989b) examines this possibility in a model with a multifactor production technology which allows for substitution between labor and three different types of capital, and nine production sectors (agriculture; mining; construction; durable goods manufacturing; nondurable goods manufacturing; transportation, communication and utilities; wholesale and retail trade; finance, insurance, and real estate; and other services). Auerbach finds that across a wide array of parameter values, the interasset distortion from nonneutral capital-income taxation are surprisingly small, but that for the Tax Reform Act of 1986, the reduction in the interasset distortion just about balanced out the increase in the tax wedge.

5.3. Equipment investment generates externalities

De Long and Summers (1991) provided evidence that suggested that economic growth is higher in the long run for countries that invest more in equipment. They argue that this pattern is inconsistent with the Solow growth model, which predicts that the level of investment should have no effect on economic growth in the long run. They argue that "learning by doing" may explain this pattern. Individuals who have learned to operate one type of machine, may have an easier time picking up the operation of a different one as well. Countries with such flexible workers may benefit in the long run.

If this correlation proved reliable, then it would provide a justification for subsidizing equipment investment. Auerbach, Hassett and Oliner (1994) provide evidence, however, that suggests that the De Long and Summers' results were consistent with the predictions of the Solow model given the time periods studied. The argument goes like this. In the very short run, a regression that relates the change in capital and labor to the change in output should produce coefficient estimates that are related to the marginal product of each input. Because equipment investment has a higher depreciation rate than structures investment, it should have a higher short-run marginal product (and regression coefficient) if the regression is run using annual data. The Solow model says that capital intensity does not matter in the long run, so the coefficients on investment in regressions using growth over, say, a century, should be zero. Auerbach, Hassett, and Oliner show that regressions like those reported in De Long and Summers will have coefficients that depend on the length of time over which growth is being measured. They find that the regressions match the pattern predicted by the Solow model very closely.

5.4. Investment incentives do not work because some firms face financing constraints

In contrast to the frictionless capital markets in the standard neoclassical model considered to this point, earlier applied research on investment, especially the work of Meyer and Kuh (1957), stressed the significance of financial considerations

(particularly internal funds or net worth) for business investment. Since the mid-1960s, however, most applied research on investment isolated "real" firm decisions from "financing". The intellectual justification for this shift in approach drew on the seminal work by Modigliani and Miller (1958), who demonstrated the irrelevance of financial structure and financial policy for real investment decisions under certain conditions. The central Modigliani–Miller result, which facilitated the early development of the neoclassical model, was that a firm's financial structure will not affect its value in frictionless capital markets. As a result, if their assumptions are satisfied, real firm decisions, motivated by the maximization of shareholders' claims, are independent of financial factors such as the availability of internal funds.

The assumption of representative firms (in terms of trade on capital markets) is common to most research programs in the neoclassical tradition. That is, the same empirical model applies to all firms. Therefore, tests could not ascertain whether the observed sensitivity of investment to financial variables differs across firms and whether these differences in sensitivity explain the weak apparent relationship between the measured user cost and investment. Contemporary empirical studies of information and incentive problems in the investment process have moved beyond the assumption of representative firms by examining firm-level panel data in which firms can be grouped into "high-net-worth" and "low-net-worth" categories. For the latter category, changes in net worth or internal funds often appear to affect investment, holding constant underlying investment opportunities (desired investment)[24]. Following Fazzari, Hubbard and Petersen (1988a), empirical researchers have placed firms into groups as *a priori* "financially constrained" or "financially unconstrained".

Two aspects of the findings of this research program are noteworthy in the context of measuring incentives to invest. First, numerous empirical studies have found that proxies for internal funds have explanatory power for investment, holding constant Q, the user cost, or accelerator variables [see the review of studies in Hubbard (1998)]. This suggests that tax policy may have effects on investment by constrained firms beyond those predicted by neoclassical approaches. (Indeed, returning to the "accelerator" analogy, Bernanke, Gertler and Gilchrist (1996) argue that this literature describes a "financial accelerator".) In particular, the quantity of internal funds available for investment is supported by the average tax on earnings from existing projects. In this sense, average as well as marginal tax rates faced by a firm affect its investment decisions.

Second, empirical studies of financing constraints generally find that the frictionless neoclassical model is rejected only for the groups of firms that *a priori* are financially constrained [see, e.g., Calomiris and Hubbard (1995) and Hubbard, Kashyap and Whited (1995)], and in most papers in the literature, this set of firms accounts for only

[24] For reviews of the theoretical literature, see Bernanke, Gertler and Gilchrist (1996) and Hubbard (1998).

a small fraction of aggregate investment. Hence, while the shadow value of internal funds may not be well captured for some firms in standard representations of the neoclassical approach, the neoclassical model with convex adjustment costs yields reasonable estimated values of marginal adjustment costs for most firms[25].

5.5. Investment incentives are reflected in prices of capital goods

One scenario under which investment incentives might have a small economic impact but at high revenue costs is when the increase in investment demand following a tax decrease is offset by an increase in prices of investment goods. Such a scenario would be important if the supply of capital goods were fixed in the short run, or at least highly inelastic. While it is implausible that the supply function for most individual capital-goods manufacturers is perfectly elastic, the effective supply of capital goods to a given domestic market might well be highly elastic in the long run if the world market for capital goods is open.

Goolsbee (1998) addresses this important issue directly, using disaggregated price and tax data to investigate the extent to which investment incentives stimulate increases in the prices of capital goods. Goolsbee finds a significant response of capital-goods prices to investment subsidies, and concludes that investment tax credits are largely captured by capital-goods manufacturers.

Using data for the United States and ten other countries, Hassett and Hubbard (1998) find that local investment tax credits have a negligible effect on prices paid for capital goods; indeed, they find that the capital-goods prices for most countries are very highly correlated, and that the movements of these over time are consistent with "the law of one price". In addition, using disaggregated data on asset-specific investment-good prices and tax variables for the United States, we find that tax parameters have no effect on capital-goods prices. The conclusion that tax policy in the United States does not affect the world price of capital goods is especially meaningful, given the relative size of the US economy. Taken together, these tests suggest that the effects of investment tax policy have not been muted in a significant way by upward-sloping supply schedules for capital goods.

Hassett and Hubbard (1998) explore the reasons for the disagreement. Goolsbee's price regressions may suffer from "spurious regression" problems, because the price series used are highly nonstationary, and not cointegrated with the tax variables. When the data are differenced to correct for these factors, Goolsbee's strong relationships disappear. However, measurement-error problems are exacerbated by differencing, and the presence or lack thereof of a US price effect is the subject of ongoing debate.

[25] Average tax rates on profits may, nonetheless, affect investment decisions of smaller or entrepreneurial firms [see Fazzari, Hubbard and Petersen (1988b), Gentry and Hubbard (2000) and Holtz-Eakin, Joulfaian and Rosen (1994)].

5.6. *Investment incentives are reflected in higher interest rates*

Recall that Cummins, Hassett and Hubbard (1994) identify high estimated elasticities of investment effectively from a substitution between tax-favored and tax-disadvantaged assets. Recall also that while this substitution takes place, time-series changes in the user cost are not obviously associated with significant changes in investment (Figure 1).

Hines (1998) argues that traditional user-cost models ignore problems of asymmetric information and bankruptcy, which generate divergent interests for bondholders and equity-holders. While equity-holders want to maximize after-tax returns, bondholders want the firm to maximize before-tax returns, thereby maximizing the value of the firm if it is in default. Bondholders recognize that an investment tax credit spurs investments that, all else being equal, reduce the pretax profitability of the firm, reducing payoffs to bondholders in the event of bankruptcy. Anticipating this possibility bondholders may demand that firms pay them higher interest rates to offset the higher risk. In Hines' model, this interest-rate response can in principle be large enough that aggregate investment does not respond to an investment incentive, even though firms substitute substantially between tax-favored and tax-disadvantaged assets. Hines shows that bond yields responded in the way predicted by his model to the Tax Reform Act of 1986. Because the Act removed an equipment subsidy, bondholders in his models should have been pleased. Indeed, at the announcement of the Act, interest rates on corporate bonds dropped by between 15 and 40 basis points, with lower-grade bond rates dropping more.

Cummins, Hassett and Hubbard (1994) argued that simultaneity problems made the identification of the user-cost elasticity impossible with time-series data alone. They argued that the large elasticities they found were consistent with large aggregate elasticities as well, and that the concurrent swings in investment and the user cost were not evident in the time-series data because of simultaneity. Hines' model offers one explanation for one channel of such simultaneity, but there are many other potential ones as well (e.g., accelerator effects). Until all of the plausible effects are identified, making precise predictions about the aggregate effects of tax reforms will be difficult.

5.7. *The economy has "too much" capital*

While it is instructive to ask how effective investment incentives are at increasing the fixed capital stock, a more important question remains: What is the social value of the increase in the fixed capital stock?

Theoretical research has demonstrated that perfectly competitive economies do not necessarily converge to the "correct" capital stock. Indeed, Diamond (1965) demonstrated that a competitive economy can reach a steady state in which there is "too much" capital, in the sense that the economy is investing more than it is earning in profit. In this case, individuals can be made better off if they are forced to consume a portion of the capital stock. When evaluating investment incentives, it is crucial for

policy analysis to evaluate whether the economy is operating with "too much" or "too little" capital.

The classic "golden-rule" literature offers benchmarks for guidance. In the approach of Phelps (1961), the golden-rule level of the capital stock relative to output is achieved when the marginal product of capital (R) net of depreciation (δ), equals the sum of the rate of growth of the labor force (n) and the rate of labor-augmenting technical change (g), or [as in Blanchard and Fischer (1989)]:

$$R = \delta + n + g.$$

Alternatively, in the optimal-growth literature, Ramsey's (1928) golden-rule levels require that the marginal product of capital net of depreciation equal the sum of the social rate of time preference (ρ) and the elasticity of marginal social utility with respect to per capita consumption (ϕ), the Ramsey golden-rule levels of capital can be less than the Phelps golden-rule levels.

Following the convention in neoclassical models of the capital stock, we assume a Cobb–Douglas technology, so that the ratio of the steady-state golden-rule capital stock (K^*) to output (Y) equals the ratio of capital's share in output (α) to the optimal gross marginal product of capital (R^*). Moreover, the golden-rule level of net investment (I^*) relative to output equals ($n+g$) times the capital–output ratio. Hence:

$$\frac{K^*}{Y} = \frac{\alpha}{R^*}, \tag{18a}$$

$$\frac{I^*}{Y} = (n+g)\frac{K^*}{Y}. \tag{18b}$$

One can account for different types of capital by noting that, in equilibrium, the net rates of return on the alternative types are equal. Hence one can substitute into Equations (18a) and (18b) measures of α_k for each type of capital and the relevant R^* and Y (given differences in depreciation), and solve for the golden-rule levels of capital stocks.

Using a range of parameter values in the golden-rule expressions in Equations (18a) and (18b), Cohen, Hassett and Kennedy (1995) compare golden-rule and actual values over the period from 1980 to 1994. Table 4, which we excerpt from several tables in that study, indicates that for benchmark parameter values, equipment investment and capital stocks are below their golden-rule levels (assuming 1980–1994 is sufficiently long to characterize a steady state), while residential investment and the residential capital stocks, which received significant tax subsidies over this time period, are near or above their golden-rule levels. Cohen, Hassett, and Kennedy also show that these conclusions are not changed if the key parameters are allowed to vary across a broad range of plausible values.

Alternatively, several authors have attempted to evaluate the optimality of the US capital stock by relating various interest rates to the growth rate of GDP in the steady state. According to the golden rule, these should equal one another. If there is

Table 4
Benchmark golden-rule and actual levels of I_{net}/Y and K/Y [a]

Type of capital	Golden-rule level (%) [b]		Actual level (%) (1980–1994 average)
	Phelps	Ramsey	
Net investment as percent of GDP			
Total fixed	8.3	6.0	4.2
Business fixed	4.8	3.6	2.4
Producers durable equipment	2.4	2.0	1.3
Nonresidential structures	2.0	1.3	1.2
Residential	2.7	1.6	1.8
Ratio of capital stock to GDP			
Total fixed	3.3	2.4	1.9
Business fixed	1.9	1.4	1.0
Producers durable equipment	1.0	0.8	0.5
Nonresidential structures	0.8	0.5	0.5
Residential	1.1	0.6	0.9

[a] Source: Cohen, Hassett and Kennedy (1995), Table 2.
[b] Benchmark parameter values are:
Labor force growth rate = 01;
Rate of labor-augmenting technical change = 0.15;
Social discount rate = 0.12;
Social intertemporal elasticity of substitution (Φ) = 3.
$\alpha_{Totalfixed} = 0.30$, $\alpha_{Businessfixed} = 0.24$, $\alpha_{Equipment} = 0.18$, $\alpha_{Structures} = 0.06$, $\alpha_{Residential} = 0.06$.

too much capital, the interest rate will be lower than the growth rate. On the one hand, Tobin (1965), Solow (1970) and Feldstein (1977) argue that the marginal productivity of capital one obtains from accounting profits estimates is about 10 percent, well above the interest rate, and at a level that suggests there is too little capital. On the other hand, Ibbotson (1998) calculates a mean real return on US Treasury bills from 1926 to 1997 of only 0.7 percent, suggesting that there may have been too much capital. Of course, the answer to the question using interest rates and stock-market returns depends critically on the relevant weights associated to each return, and on the impact of risk, difficult questions that suggest that this approach may not be likely to lead to decisive conclusions.

Abel, Mankiw, Summers and Zeckhauser (1989) pursue an alternative strategy for evaluating whether the US capital stock is greater or less than the optimal level. In a stochastic setting with a very general production technology, they demonstrate that an economy is dynamically inefficient if it invests more than the returns from capital. They show that the economy is dynamically efficient – and hence in the range in

which stimulative tax policy might have positive social returns – if the returns from capital exceed investment. Using their terminology, the key question is whether the capital stock is a "sink" or a "spout", that is, whether the capital sector produces cash flow, or consumes it. This observation is a useful contribution because it allows one to base judgment about dynamic efficiency on readily observable cash flows. Abel et al. conclude that the economy is dynamically efficient. Thus, both capital-stock data and "cash-flow" data suggest that, by raising the stock of equipment capital, investment incentives may have positive social returns[26].

6. Applications to other public policies toward investment

The finding of significant short-term and long-term effects of tax-related neoclassical fundamentals on equipment investment suggests applications to current policy debates. In particular, we evaluate in this section consequences for the user cost and investment of a reduction in inflation and a switch from an income tax to a broad-based consumption tax.

6.1. Low inflation as an investment subsidy

Many economists [see, e.g., Feldstein (1976) and King and Fullerton (1984)] have argued that under fairly general assumptions, a reduction in the rate of inflation provides a relatively costless stimulus to business fixed investment by reducing the user cost of capital. Returning to the expression for the user cost, there are two channels through which expected inflation affects investment decisions. First, for given values of r and δ, the user cost varies positively with the level of expected inflation π because the present value of depreciation allowances – which is formed using the nominal rate, $\rho+\pi$, as a discount rate – varies inversely with inflation owing to historical-cost depreciation. Second, inflation affects the real cost of funds, ρ. In this section, we briefly illustrate this second channel and calculate the extent to which lower inflation over the past decade led to a reduction in the user cost of capital.

In a small open economy, the real cost of debt would be determined in world capital markets and would be exogenously given to firms. If the capital market were closed, the marginal tax rate of the holder of debt would affect the interest rates that firms pay. That is, local debt holders require a fixed real after-tax return, r:

$$r = i(1 - t_p) - \pi,$$

where i is the nominal interest rate on corporate debt, t_p is the marginal personal tax rate on interest income, and π is the expected rate of inflation. The inflation-

[26] A note of caution is in order, however, because the golden-rule models are developed for a closed economy: it is difficult to extend the comparison to domestic versus foreign fixed capital.

premium component of interest income is taxable to bondholders. The firm's real cost of debt,

$$\rho_d = i(1 - t_c) - \pi, \tag{19}$$

depends on its own marginal income tax rate, t_c, because under current US tax law, nominal interest payments on corporate debt are fully deductible. Combining the two previous expressions relates the firm's real cost of debt to the investor's required return and marginal tax rate:

$$\rho_d = (r + \pi)\frac{(1 - t_c)}{(1 - t_p)} - \pi. \tag{20}$$

Note that for a given r, inflation has very little effect on the cost of debt financing, if – as is likely the case in the United States – t_c is approximately equal to t_p. In this case, the impact of lower inflation on the cost of debt financing depends crucially on the assumption that the marginal debtholder is taxable. If the marginal debtholder is a pension fund (whose income is not taxed under current law), then lower inflation unambiguously increases the cost of debt financing. Firms receive smaller interest deductions, and pension funds do not accrue an offsetting reduction in tax liability. Alternatively,

$$\rho_d = \frac{r(1 - t_c) + (t_p - t_c)\pi}{(1 - t_p)}.$$

The firm's real cost of equity finance, ρ_e, is defined as

$$\rho_e = d + e - \pi, \tag{21}$$

where d is the dividend–price ratio and e is investor's required ex-dividend nominal return to equity. In what follows, we continue to adopt the tax-capitalization view of equity taxation [see Auerbach (1979), Bradford (1981) and King (1977)], which suggests that the relevant equity tax rate is the effective capital-gains rate, regardless of dividend policy. This view is premised on the assumptions that equity funds come primarily from retained earnings (i.e., lower dividends paid out of current earnings) rather than from new share issues, and that earnings distributions to shareholders are primarily through dividends rather than share repurchases. The idea is that taxes on dividend distributions are capitalized into the value of the equity rather than imposing a burden on the returns to new investment, as would be the case if new investment were financed by the issue of new shares.

Under the tax-capitalization view, marginal equity funds for a dividend-paying firm are provided by retained earnings. Hence the opportunity cost to the shareholder of a dollar of new investment is reduced by the dividend taxes foregone (evaluated at the dividend tax rate τ_d), net of the increased tax burden on the capital gains induced by the

accrual (evaluated at the accrual-equivalent tax rate on capital gains, t_c). Because the value of new investment per dollar invested, q, should equal its cost to the shareholder, the equilibrium cost of retaining a dollar is $q = 1 - t_p + t_g q$, which implies that $q = (1 - t_p)/(1 - t_g)$.

Capital-market equilibrium requires additionally that the after-tax rate of return on the firm's investment in (nominal terms) equals the investor's required rate of return, ρ_i. Following Auerbach (1983), for a given value of q:

$$\tilde{\rho}_i = (1 - \tau_d)\frac{d}{q} + (1 - c)\,e,$$

where c is the accrual-equivalent capital gains tax rate. Substituting for q and converting to real terms:

$$\rho_i = \tilde{\rho}_i - \pi = (1 - t_g)(d + e) - \pi,$$

where the subscript i refers to the marginal investor. Combining terms, we can express the firm's real cost of equity financing as

$$\rho_e = \frac{\rho_i}{(1 - t_g)} + \frac{t_g}{(1 - t_g)}\pi. \tag{22}$$

Further, in equilibrium, investors' after-tax real returns on debt and equity, adjusted for a risk premium, X, must be equal, i.e., $r = \rho_i + X$. Solving for ρ_i and substituting the resulting expression into Equation (22), we get:

$$\rho_e = \frac{X}{(1 - t_g)} + \frac{(1 - t_p)}{(1 - t_g)}R - \pi. \tag{23}$$

Differentiation of this expression, assuming that the risk premium is unaffected by inflation, and deferring consideration of the dividend term to below, we find that, for a given r (i.e., in the tax-adjusted Fisher-effect case), lower inflation unambiguously *reduces* the cost of equity finance by the factor $t_g/(1 - t_g)$. This term captures the "inflation tax" paid by shareholders who receive purely nominal gains; taxation of real capital gains would eliminate this effect. There is another, offsetting effect however, if the traditional Fisher effect holds (in which the nominal bond rate rises point-for-point with inflation). In this case, lower inflation also raises r by t_p times the change in inflation and, hence, ρ_i by the same amount. As a result, the total impact on the firm's real cost of equity financing in this case depends on the difference between the personal tax rate on interest and the effective capital-gains tax rate.

Cohen, Hassett and Hubbard (1999) calculate the marginal effects on the user cost of lowering inflation[27], and explore the effects of differing assumptions concerning

[27] Earlier empirical studies of the effect of inflation on real business tax burdens include Feldstein and Summers (1979) and Auerbach (1983). Cohen, Hassett and Hubbard also allow for inflation to increase taxes because of effects of inflation on the cost of carrying inventories.

these effects on their conclusion. (In addition to the effect cited above, they also control for the fact that inflation changes the present value of depreciation deductions). They estimate that the current value of the user cost for equipment investment is about 0.22, and they conclude that a one-percentage-point permanent decrease in inflation lowers the user cost by about one-half a percentage point, assuming that the after-tax Fisher effect holds. In their calculations, the incremental effect of each additional percentage-point reduction in inflation is approximately the same. Thus, if the annual inflation rate were reduced from four percent to zero, the user cost of capital would decline about two percentage points – proportionally by about ten percent. On the one hand, given the elasticity estimates reviewed earlier, this "tax cut" would provide a significant stimulus to investment. On the other hand, if the pure Fisher effect holds, then the stimulus of lower inflation would be very small.

6.2. Moving from an income tax to a consumption tax

Under the income tax, the user cost of capital is influenced by the corporate tax rate, investment tax credits, and the present value of depreciation allowances. Under a broad-based consumption tax, firms pay tax on the difference between receipts and purchases from other firms. That is, there is no investment tax credit, and investment is expensed. In this case (assuming that the corporate tax rate does not change over time), the user cost of capital no longer depends on taxes. That is, under a consumption tax, taxes do not distort business investment decisions; investment decisions are based solely on non-tax fundamentals. Because US tax policy currently increases the user cost, the switch to the consumption tax lowers the user cost and increases investment.

Of course, other aggregate variables are also likely to change in response to such a large change to the tax code. For example, nominal interest rates and the supply of savings are likely to change. While it is difficult to say how large the net stimulus to investment would be, the consensus of the recent investment literature suggests that the partial-equilibrium impact on investment may be quite large.

6.3. Temporary investment incentives?

We have focused our attention on permanent changes in investment incentives. Even a casual observation of the history of investment incentives since the 1950s suggests the usefulness of analyzing consequences of temporary investment incentives. Since 1962, the mean duration of a typical state in which an ITC is in effect has been about three and one-half years, and the mean duration of the "no-ITC" state has been about the same length. Most recently, President Bush advocated a modified ITC, known as the "Investment Tax Allowance" in 1992, and President Clinton proposed an incremental ITC in early 1993; neither of these measures was enacted. What is the likely impact on aggregate capital accumulation of temporary investment incentives?

Temporary investment incentives can have even larger short-run impacts on investment than permanent investment incentives [see, e.g., Auerbach (1989a)].

Consider, for example, a temporary ITC known to last for one period. The ITC lowers the current user cost both through its effect on the price of purchasing a machine today and through the consequences of its removal tomorrow. Firms face an incentive to acquire capital goods before the credit is removed.

The large potential effects of temporary tax incentives on investment do not imply that such incentives are desirable – even if one believes that long-run investment incentives are sound tax policy. In the presence of uncertainty and adjustment costs, there is little reason to believe that policymakers can "time" investment incentives for the purposes of stabilization policy. Moreover, the use of temporary incentives increases uncertainty in business capital budgeting, making it more difficult for firms to forecast the path of the user cost of capital.

What if firms do not know the exact timing of changes in investment incentives – that is, if tax policy is uncertain? There is a substantial literature evaluating the effects of price uncertainty on investment, and the lesson from this literature is that the sign of the effect of uncertainty on investment depends crucially on assumptions about adjustment costs and returns to scale. Hartman (1972) shows that uncertainty generally increases investment in a model with constant returns and convex adjustment costs; Abel (1983) derives a similar result in continuous time. Pindyck (1988), however, shows that uncertainty can significantly lower capital formation if investment is irreversible and if returns to scale are decreasing. We described Pindyck's intuition earlier: In an uncertain world, there is a gain to delaying investment – the option value of waiting – and these gains are higher the higher is the variance in the output price.

It is important to note, however, that tax-policy uncertainty is not the same as price uncertainty. Models of uncertainty often assume that the price follows a continuous-time random walk (Brownian motion or geometric Brownian motion). When prices follow a random walk, the appropriate rational-expectations forecast for the price at any time in the future is today's price, and the future path of the price is unbounded. Unlike most prices, however, tax parameters tend to remain constant for a period of time, and then jump to new values. In addition, investment incentives tend to be mean-reverting: When an investment incentive is high, it is likely to be reduced in the future; when an investment incentive is low, it is likely to be increased in the future. With these properties, the normal gain to waiting in a model with irreversibility is reduced significantly when an investment tax credit is "on": Because the firm fears that the credit might be eliminated, it is more likely to invest today while the credit is still effective. Indeed, this effect dominates the reverse effect in the state in which there is no investment tax credit, so that increasing tax-policy uncertainty can raise aggregate investment; [see Hassett and Metcalf (1999) and Alvarez, Kanniainen and Södersten (1998)].

As with the case of temporary investment incentives generally, this result does not imply that random tax policy is desirable. Most existing studies analyze investment in a partial-equilibrium setting wherein there are no utility costs to bunching capital formation. In a general-equilibrium setting, Bizer and Judd (1989) show that welfare is reduced significantly by random investment-tax policy. The randomness has

a negative impact because consumers wish to smooth consumption, and fluctuations in investment credits make smoothing costly.

7. Conclusions

Simple theoretical models of responses of investment dynamics to tax variables suggest important effects of personal and business taxation on investment and the long-run capital stock. The empirical study of business investment has gone very far, very fast. Ten years ago, almost no economist believed that the investment demand elasticity was much different from zero; in a recent survey of specialists in labor and public economics [Fuchs, Krueger and Poterba (1998)], the median respondent indicated that a decline in the user cost from a switch to expensing would increase investment an amount consistent with an elasticity of about unity. Perhaps this agreement reflects the strong biases of economists, but recent empirical research is consistent with this broad agreement. A consensus has emerged that investment demand is sensitive to taxation and neoclassical investment models are useful for policy analysis.

While there is a consensus about the nature and magnitude of tax policy on investment demand, considerable uncertainty remains regarding the structure of adjustment costs and the short-run dynamic effects of tax reforms. Consistent with our analysis of equilibrium investment outcomes, ascertaining the effects of tax policy on equilibrium investment requires additional research to examine responsiveness of interest rates, output, and the stock market to tax-policy changes.

References

Abel, A.B. (1980), "Empirical investment equations: an integrative framework", Journal of Monetary Economics 12:39–91.

Abel, A.B. (1983), "Optimal investment under uncertainty", American Economic Review 73:228–233.

Abel, A.B. (1990), "Consumption and investment", in: Benjamin M. Friedman and Frank H. Hahn, eds., Handbook of Monetary Economics, Vol. 2 (North-Holland, Amsterdam) pp. 725–778.

Abel, A.B., and O.J. Blanchard (1986), "The present value of profits and cyclical movements in investment", Econometrica 54:249–273.

Abel, A.B., and J.C. Eberly (1994), "A unified model of investment under uncertainty", American Economic Review 84:1369–1384.

Abel, A.B., and J.C. Eberly (1996a), "Optimal investment with costly irreversibility", Review of Economic Studies 3:581–593.

Abel, A.B., and J.C. Eberly (1996b), "Investment and q with fixed costs: an empirical analysis", Mimeograph (The Wharton School, Pennsylvania).

Abel, A.B., and J.C. Eberly (1999), "The effects of irreversibility and uncertainty on capital accumulation", Journal of Monetary Economics 44:339–371.

Abel, A.B., N.G. Mankiw, L.H. Summers and R.J. Zeckhauser (1989), "Assessing dynamic efficiency: theory and evidence", Review of Economic Studies 56:1–20.

Aftalian, A. (1909), "La reálité des surproductions générales, essui d'une théorie des crises générales et periodiques", Revue d'Economie Politique 1909.

Alvarez, L., V. Kanniainen and J. Södersten (1998), "Tax policy uncertainty and corporate investment: a theory of tax-induced investment spurts", 1(7):17–48.
Auerbach, A.J. (1979), "Wealth maximization and the cost of capital", Quarterly Journal of Economics 93:433–446.
Auerbach, A.J. (1983), "Taxation, corporate financial policy, and the cost of capital", Journal of Economic Literature 21:905–940.
Auerbach, A.J. (1989a), "Tax reform and adjustment costs: the impact on investment and market value", International Economic Review 30:939–962.
Auerbach, A.J. (1989b), "The deadweight loss from 'non-neutral' capital income taxation", Journal of Public Economics 40:1–36.
Auerbach, A.J., and K.A. Hassett (1991), "Recent U.S. investment behavior and the Tax Reform Act of 1986: a disaggregate view", Carnegie-Rochester Conference Series on Public Policy 35:185–215.
Auerbach, A.J., K.A. Hassett and S. Oliner (1994), "Reassessing the social returns to equipment investment", Quarterly Journal of Economics 109:789–802.
Barnett, S.A., and P. Sakellaris (1998), "Nonlinear response of firm investment to Q: testing a model of convex and nonconvex adjustment costs", Journal of Monetary Economics (October 1998):261–288.
Bernanke, B., H. Bohn and P.C. Reiss (1988), "Alternative nonnested specification tests of time-series investment models", Journal of Econometrics 37:293–326.
Bernanke, B., M. Gertler and S. Gilchrist (1996), "The financial accelerator and the flight to quality", Review of Economics and Statistics 78:1–15.
Bertola, G., and R.J. Caballero (1990), "Kinked adjustment costs and aggregate dynamics", in: Olivier J. Blanchard and Stanley Fischer, eds., NBER Macroeconomics Annual (MIT Press, Cambridge).
Bizer, D., and K. Judd (1989), "Taxation and uncertainty", American Economic Review 79:331–336.
Blanchard, O.J., and S. Fischer (1989), Lectures on Macroeconomics (MIT Press, Cambridge).
Bond, S., and C. Meghir (1994), "Dynamic investment models and the firm's financial policy", Review of Economic Studies 61:197–222.
Bosworth, B.P. (1985), "Taxes and the investment recovery", Brookings Papers on Economic Activity 1:1–38.
Bradford, D. (1981), "The incidence and allocation effects of a tax on corporate distribution", 15(1):1–22.
Caballero, R.J. (1994), "Small sample bias and adjustment costs", Review of Economics and Statistics 76(1):52–58.
Caballero, R.J. (1999), "Aggregate investment", in: John B. Taylor and Michael Woodford, eds., Handbook of Macroeconomics (North Holland, Amsterdam).
Caballero, R.J., E.M.R.A. Engel and J.C. Haltiwanger (1995), "Plant-level adjustment and aggregate investment dynamics", Brookings Papers on Economic Activity 2:1–54.
Calomiris, C.W., and R.G. Hubbard (1995), "Internal finance and investment: evidence from the undistributed profits tax of 1937–1938", Journal of Business 68:443–482.
Carroll, R., D. Holtz-Eakin, M. Rider and H.S. Rosen (2000), "Entrepreneurs, income taxes, and investment", in: Joel Slemrod, ed., Does Atlas Shrug? The Economic Consequences of Taxing the Rich (MIT Press, Cambridge) pp. 427–455.
Chirinko, R.S. (1987), "The ineffectiveness of effective tax rates on business investment: a critique of Feldstein's Fisher–Schultz Lecture", Journal of Public Economics 32:369–387.
Chirinko, R.S. (1993), "Business fixed investment spending: modeling strategies, empirical results, and policy implications", Journal of Economic Literature 31:1875–1911.
Chirinko, R.S., and R. Eisner (1983), "Tax policy and in major macroeconomic models", Journal of Public Economics 20:139–166.
Chirinko, R.S., S.M. Fazzari and A.P. Meyer (1999), "How responsive is business capital formation to its user cost? An explanation with micro data," Journal of Public Economics 74(1)53–80.
Clark, J.M. (1917), "Business acceleration and the law of demand", Journal of Political Economy 25:217–235.

Clark, P.K. (1993), "Tax incentives and equipment investment", Brookings Papers on Economic Activity 1:317–339.
Cochrane, J. (1991), "Production-based asset pricing and the link between stock returns and economic fluctuations", Journal of Finance 46:209–237.
Cohen, D., K.A. Hassett and J. Kennedy (1995), "Are U.S. investment and capital stocks at their optimal levels?", FEDS Working Paper 9532 (Board of Governors of the Federal Reserve System).
Cohen, D., K.A. Hassett and R.G. Hubbard (1999), "Inflation and the user cost of capital: does inflation still matter?", in: Martin Feldstein, ed., Costs and Benefit of Price Stability (University of Chicago Press, Chicago) pp. 199–230.
Cummins, J.G., and M. Dey (1998), "Taxation, investment, and firm growth with heterogeneous capital", Mimeograph (New York University).
Cummins, J.G., and R.G. Hubbard (1995), "The tax sensitivity of foreign direct investment: evidence from firm-level panel data", in: Martin Feldstein, James R. Hines Jr, and R. Glenn Hubbard, eds., The Effects of Taxation on Multinational Corporations (University of Chicago Press, Chicago) pp. 123–147.
Cummins, J.G., K.A. Hassett and R.G. Hubbard (1994), "A reconsideration of investment behavior using tax reforms as natural experiments", Brookings Papers on Economic Activity 2:1–74.
Cummins, J.G., K.A. Hassett and R.G. Hubbard (1995a), "Have tax reforms affected investment?", in: James M. Poterba, ed., Tax Policy and the Economy, Vol. 9 (MIT Press, Cambridge) pp. 131–149.
Cummins, J.G., T.S. Harris and K.A. Hassett (1995b), "Accounting standards, information flow, and firm investment behavior", in: Martin Feldstein, James R. Hines and R. Glenn Hubbard, eds., The Effects of Taxation on Multinational Corporations (University of Chicago Press, Chicago) pp. 181–221.
Cummins, J.G., K.A. Hassett and R.G. Hubbard (1996), "Tax reforms and investment: a cross-country comparison", Journal of Public Economics 62:237–273.
De Long, J.B., and L.H. Summers (1991), "Equipment spending and economic growth", Quarterly Journal of Economics 106:445–502.
Diamond, P.A. (1965), "National debt in a neoclassical growth model", American Economic Review 55:1126–1150.
Dixit, A.K., and R.S. Pindyck (1994), Investment Under Uncertainty (Princeton University Press, Princeton).
Doms, M., and T. Dunne (1994), "Capital adjustment patterns in manufacturing plants", Discussion Paper 94-11 (Center for Economic Studies, U.S. Bureau of the Census, Washington, D.C.).
Doyle, J., and T.M. Whited (1998), "Fixed costs of adjustment, coordination, and industry investment", Mimeograph (James Madison University).
Eisner, R. (1969), "Tax policy and investment behavior: comment", American Economic Review 59:379–388.
Eisner, R. (1970), "Tax policy and investment behavior: further comment", American Economic Review 60:746–752.
Eisner, R., and M.I. Nadiri (1968), "Investment behavior and neoclassical theory", Review of Economics and Statistics 50:369–382.
Eisner, R., and R.H. Strotz (1963), "Determinants of business investment", in: Impacts of Monetary Policy, studies prepared for the Commission on Money and Credit (Prentice-Hall, Englewood Cliffs, N.J.).
Fazzari, S.M., R.G. Hubbard and B.C. Petersen (1988a), "Financing constraints and corporate investment", Brookings Papers on Economic Activity 1:141–195.
Fazzari, S.M., R.G. Hubbard and B.C. Petersen (1988b), "Investment, financing decisions, and tax policy", American Economic Review 78:200–205.
Feldstein, M. (1976), "Inflation, income taxes, and the rate of interest: a theoretical analysis", American Economic Review 66:809–820.
Feldstein, M. (1977), "Does the United States save too little?", American Economic Review 67:116–121.
Feldstein, M. (1982), "Inflation, tax rules, and investment: some econometric evidence." Econometrica 50:825–862.

Feldstein, M., and L.H. Summers (1979), "Inflation and the taxation of capital income in the corporate sector", National Tax Journal 32:445–470.
Fisher, I. (1930), The Theory of Interest (Macmillan, New York).
Fuchs, V.R., A.B. Krueger and J.M. Poterba (1998), "Economists views about parameters, values, and policies: survey results in labor and public economics", Journal of Economic Literature 36:1387–1425.
Gentry, W., and R.G. Hubbard (2000), "Entrepreneurship and household saving", Working Paper (National Bureau of Economic Research).
Gilchrist, S. (1991), "An empirical analysis of corporate investment and financing hierarchies using firm-level panel data", Mimeograph (Board of Governors of the Federal Reserve System).
Gilchrist, S., and C.P. Himmelberg (1995), "Evidence on the role of cash flow in reduced-form investment equations", Journal of Monetary Economics 36:541–572.
Gilchrist, S., and C.P. Himmelberg (1998), "Investment fundamentals and finance", in: Ben Bernake and Julio J. Rotemberg, eds., NBER Macroeconomics Annual (MIT Press, Cambridge) pp. 223–262.
Goolsbee, A. (1998), "Investment tax incentives and the price of capital goods", Quarterly Journal of Economics 113:121–148.
Goolsbee, A. (2000a), "Measurement error and the cost of capital", National Tax Journal 53:215–228.
Goolsbee, A. (2000b), "Taxes and the quality of capital", Mimeograph (University of Chicago).
Goolsbee, A., and D.B. Gross (1997), "Estimating adjustment costs with data on heterogenous capital goods", Working Paper 6342 (National Bureau of Economic Research).
Gould, J.P. (1968), "Adjustment costs in the theory of investment of the firm", Review of Economic Studies 35:47–55.
Griliches, Z., and J.A. Hausman (1986), "Errors in variables in panel data", Journal of Econometrics 31:141–154.
Haavelmo, T. (1960), A Study in the Theory of Investment (University of Chicago Press, Chicago).
Hall, R.E., and D.W. Jorgenson (1967), "Tax policy and investment behavior", American Economic Review 57:391–414.
Hartman, R. (1972), "The effects of price and cost uncertainty on investment", Journal of Economic Theory 5:258–266.
Hassett, K.A., and R.G. Hubbard (1998), "Are investment incentives blunted by changes in the price of capital goods?" International Finance 1:103–126.
Hassett, K.A., and R.G. Hubbard (1999), "Hurdle rates", Mimeograph (Columbia University).
Hassett, K.A., and G.E. Metcalf (1999), "Investment with uncertain tax policy: does random tax policy discourage investment?" The Economic Journal 109(457):372–393.
Hayashi, F. (1982), "Tobin's marginal q and average q: a neoclassical interpretation", Econometrica 50:213–224.
Hines Jr, J.R. (1998), "Is it investment ramifications of distortionary tax subsidies." Working Paper 6615 (National Bureau of Economic Research).
Holtz-Eakin, D., D. Joulfaian and H.S. Rosen (1994), "Sticking it out: entrepreneurial survival and liquidity constraints", Journal of Political Economy 102:53–75.
Hubbard, R.G. (1994), "Investment under uncertainty: keeping one's options open", Journal of Economic Literature 32:1816–1831.
Hubbard, R.G. (1998), "Capital-market imperfections and investment", Mimeograph (Columbia University); Journal of Economic Literature 36:193–225.
Hubbard, R.G., and A.K. Kashyap (1992), "Internal net worth and the investment process: an application to U.S. agriculture", Journal of Political Economy 100:506–534.
Hubbard, R.G., A.K. Kashyap and T.M. Whited (1995), "Internal finance and firm investment", Journal of Money, Credit, and Banking 27:683–701.
Ibbotson, R.G. (1998), Stocks, Bonds, Bills, and Inflation: Market Results for 1926–1997 (Ibbotson and Associates, Illinois).
Jorgenson, D.W. (1963), "Capital theory and investment behavior", American Economic Review 53: 247–259.

Jorgenson, D.W. (1967), "Theory of investment behavior", in: Robert Ferber, ed., Determinants of Investment Behavior (Columbia University Press, New York).
Judd, K.L. (1985), "Short-run analysis of fiscal policy in a perfect-foresight model", Journal of Political Economy 93:298–319.
Judd, K.L. (1997), "The optimal tax rate for capital income is negative", Working Paper 6004 (National Bureau of Economic Research).
Judd, K.L. (2001), "The impact of tax reform in modern dynamic economics", in: K.A. Hassett and R.G. Hubbard, eds., Transitional Costs of Fundamental Tax Reform (AEI Press, Washington, D.C.) pp. 5–53.
King, M.A. (1977), Public Policy and the Corporation (Chapman and Hall, London).
King, M.A., and D. Fullerton, eds (1984), The Taxation of Income from Capital: A Comparative Study of the United States, United Kingdom, Sweden, and West Germany (University of Chicago Press, Chicago).
Lucas Jr, R.E. (1967), "Adjustment costs and the theory of supply", Journal of Political Economy 75:321–334.
Lucas Jr, R.E. (1976), "Econometric policy evaluation: a critique", in: Karl Brunner and Allan Meltzer, eds., The Phillips Curve and Labor Markets (Carnegie-Rochester Conference Series on Public Policy 1) pp. 19–46.
Messere, K. (1993), Tax Policy in OECD Countries: Choices and Conflicts (IBFD Publications, Amsterdam).
Meyer, J.R., and E. Kuh (1957), The Investment Decision (Harvard University Press, Cambridge).
Modigliani, F., and M.H. Miller (1958), "The cost of capital, corporation finance and the theory of investment", American Economic Review 48:261–297.
Phelps, E.S. (1961), "The golden rule of accumulation: a fable for growth men", American Economic Review 51:638–643.
Pindyck, R.S. (1988), "Irreversible investment capacity choice, and the value of the firm", American Economic Review 78:969–985.
Pindyck, R.S. (1991), "Irreversibility, uncertainty, and investment", Journal of Economic Literature 29:1110–1148.
Pindyck, R.S., and J.J. Rotemberg (1983), "Dynamic factor demands under rational expectations", Scandinavian Journal of Economics 85:223–238.
Poterba, J.M., and L.H. Summers (1985), "The economic effects of dividend taxation", in: Edward I. Ahman and Marti G. Subrahmanyan, eds., Recent Advances in Corporate Finance (Irwin, Homewood, IL).
Ramsey, F.P. (1928), "A mathematical theory of saving", Economic Journal 62:543–559.
Romer, D. (1996), Advanced Macroeconomics, (McGraw-Hill Companies, New York).
Rothschild, M. (1971), "On the costs of adjusting the capital stock", Quarterly Journal of Economics 85:605–622.
Salinger, M.A., and L.H. Summers (1983), "Tax reform and corporate investment: a microeconomic simulation study." in: Martin Feldstein, ed., Behavioral Simulation Methods in Tax Policy Analysis (University of Chicago Press, Chicago).
Shapiro, M.D. (1986), "The dynamic demand for capital and labor", Quarterly Journal of Economics 101:513–547.
Sinn, H.-W. (1987), Capital Income Taxation as Resource Allocation (North-Holland, Amsterdam).
Solow, R. (1970), Growth Theory (Oxford University Press, Oxford).
Stock, J.H., and M.W. Watson (1993), "A simple MLE of cointegrating vectors in higher order integrated systems", Econometrica 61:111–152.
Summers, L.H. (1981), "Taxation and corporate investment: a q-theory approach", Brookings Papers on Economic Activity 1:67–127.
Summers, L.H. (1987), "Investment incentives and the discounting of depreciation allowances", in:

Martin Feldstein, ed., The Effects of Taxation on Capital Accumulation (University of Chicago Press, Chicago) pp. 295–304.
Tobin, J. (1965), "Economic growth as an objective of government policy", in: Essays in Economics, Vol. 1, Macroeconomics (North-Holland, Amsterdam) pp. 174–194.
Tobin, J. (1969), "A general equilibrium approach to monetary theory", Journal of Money, Credit, and Banking 1:15–29.
Treadway, A. (1970), "Adjustment cost and variable imports in the theory of the competitive firm", Journal of Economic Theory 2:329–347.
Turnovsky, S.J. (1995), Methods of Macroeconomic Dynamics (MIT Press, Cambridge).
Uzawa, H. (1969), "Time preference and the Penrose effect in a two-class model of economic growth", Journal of Political Economy 77:628–652.
Whited, T.M. (1992), "Debt, liquidity constraints, and corporate investment", Journal of Finance 47:1425–1460.

Chapter 21

TAXATION AND ECONOMIC EFFICIENCY*

ALAN J. AUERBACH
University of California, Berkeley and NBER

JAMES R. HINES Jr.
University of Michigan and NBER

Contents

* We thank Charles Blackorby, Peter Diamond, Kenneth Judd, Louis Kaplow, Gareth Myles, Michel Strawczynski and Ronald Wendner for helpful comments on a previous draft.

Handbook of Public Economics, Volume 3, Edited by A.J. Auerbach and M. Feldstein

Abstract

This chapter analyzes the distortions created by taxation and the features of tax systems that minimize such distortions (subject to achieving other government objectives). It starts with a review of the theory and practice of deadweight loss measurement, followed by characterizations of optimal commodity taxation and optimal linear and nonlinear income taxation. The framework is then extended to a variety of settings, initially consisting of optimal taxation in the presence of externalities or public goods. The optimal tax analysis is subsequently applied to situations in which product markets are imperfectly competitive. This is followed by consideration of the features of optimal intertemporal taxation. The purpose of the chapter is not only to provide an up-to-date review and analysis of the optimal taxation literature, but also to identify important cross-cutting themes within that literature.

Keywords

deadweight loss, excess burden, optimal taxation, marginal cost of public funds, Ramsey taxation

JEL classification: H21

1. Introduction

This chapter considers a subject at the very center of public finance analysis, the distortions introduced (and corrected) by taxation. Tax-induced reductions in economic efficiency are known as *deadweight losses* or the *excess burdens* of taxation, the latter signifying the added cost to taxpayers and society of raising revenue through taxes that distort economic decisions.

Taxes almost invariably have excess burdens because tax obligations are functions of individual behavior. The alternative, pure *lump-sum* taxes, are attractive from an efficiency perspective, but are of limited usefulness precisely because they do not vary with indicators of ability to pay, such as income or consumption, that are functions of taxpayer decisions. Thus, even though tax analysis often starts with the simple case of a representative household, it is household heterogeneity and the inability fully to observe individual differences that justify the restrictions commonly imposed on the set of tax instruments. Designing an *optimal tax* system means keeping tax distortions to a minimum, subject to restrictions introduced by the need to raise revenue and maintain an equitable tax burden.

The following sections discuss the theory and measurement of excess burden and the design of optimal tax systems. The analysis draws heavily on the chapters by Auerbach (1985) and Stiglitz (1987) in the original volumes of this Handbook, interweaving the most important results contained in these two chapters with the additional insights and areas of inquiry that have appeared since their publication. For more detailed analysis and a treatment of many other topics in this literature, the reader is referred to these original essays.

1.1. Outline of the chapter

The chapter begins with the basics and then turns to selected topics. Sections 2, 3, and 4 lay out the theory of excess burden, optimal commodity taxation, and optimal income taxation. Section 5 considers the provision of public goods and the correction of externalities, and how these problems interact with the manner in which revenues are raised. Section 6 discusses the impact on tax design of deviations from perfect competition, and Section 7 extends the theory of tax design to address issues that arise in intertemporal settings. Section 8 offers some brief conclusions regarding the evolution of the literature and promising directions for future research.

2. The theory of excess burden

2.1. Basic definitions

Excess burden (or deadweight loss) is well defined only in the context of a specific comparison, or conceptual experiment. If one simply seeks "the" excess burden of a

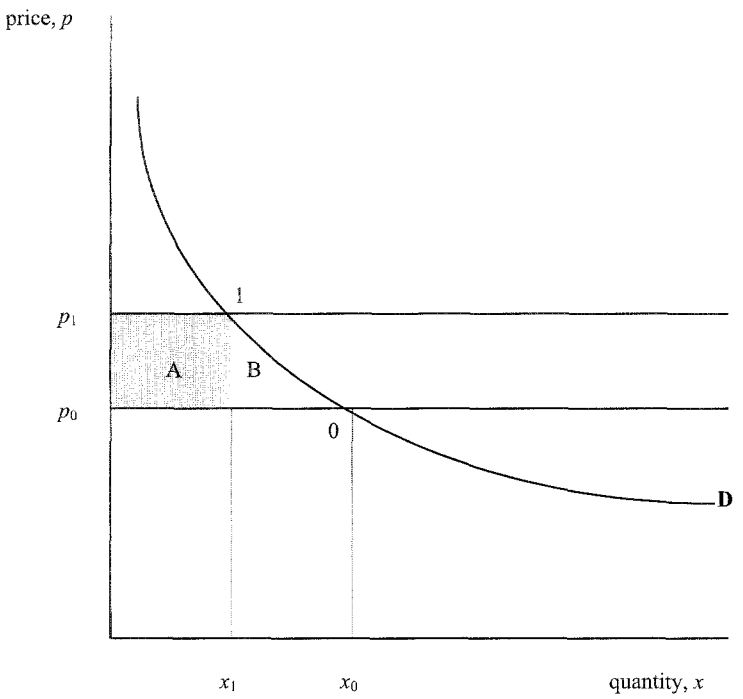


Fig. 1. The measurement of excess burden

particular tax policy, there are many equally plausible answers, so in order to obtain a unique meaning, it is necessary to be more specific. For example, the excess burden of a 10 percent tax on retail sales varies not only with the initial conditions of the tax system, but also with the direction of change, i.e., whether the tax is being added or removed.

To illustrate this ambiguity and its resolution, consider the simple case in which there are two goods, an untaxed numeraire good and a second good with a constant relative producer price of p_0. In the absence of taxation, a population of identical consumers [1] demands quantity x_0 of the second good, as depicted by point 0 in Figure 1. The imposition of a tax per unit of $p_1 - p_0$ raises the consumer price of the taxed good to p_1, with the producer price remaining at p_0. Thus, the quantity purchased falls to x_1, and the government collects revenue equal to $(p_1 - p_0)x_1$, as represented in the figure by the shaded area labeled A.

[1] We limit our discussion of excess burden to the case of identical consumers, thereby sidestepping issues of aggregation that arise in the case of heterogeneous consumers. See Auerbach (1985) for further discussion.

What is the excess burden of this tax? If one were to use the Marshallian measure of the consumers' surplus generated by consumption in this market – the area under the demand curve, D, between $x=0$ and $x=x_0$ – it would appear that consumers lose an area equal to that of regions A+B, or B in excess of the revenue actually collected. By this approach, the roughly triangular area B – commonly known as a "Harberger" triangle in recognition of Arnold Harberger's influential empirical contributions – measures the excess burden of the tax.

Unfortunately [see Auerbach (1985)], this particular measure of excess burden is not uniquely defined in a setting with more than one tax, due to the well-known problem of *path dependence* of consumers' surplus: the measure of excess burden is affected by the order in which one envisions the taxes being imposed. Path dependence is disconcerting, but more importantly reflects the imprecision of consumers' surplus-based measures of excess burden. There is no well-defined economic question to which the difference between the change in consumers' surplus and tax revenue is the answer. Thus, economists have sought alternative measures of excess burden that are not path-dependent and that answer meaningful questions.

Path dependence does not arise if excess burden is measured by Hicksian consumers' surplus, based on schedules that hold utility, rather than income, constant as prices vary. Because actual tax-policy changes typically do not hold utility constant, it is therefore necessary to construct a measure based on a conceptual experiment in which utility is held constant. One intuitive experiment is to imagine that, as a tax is imposed, utility is held constant at its pre-tax level. Graphically, in Figure 2, this measure is based on the compensated demand curve $D^c(u_0)$, which by definition passes through the original, no-tax equilibrium point 0. If the tax is imposed, and consumers are compensated to remain at original utility levels, then demand follows this schedule and the tax reduces consumption to point 1′. At this point, revenue raised is the sum of areas A and C, rather than the actual level of revenue represented by area A, because compensation induces consumers to purchase more of the taxed good (if, as is assumed here, the good is normal) and hence pay more taxes. Excess burden is defined as the amount, in excess of this revenue, that the government must compensate consumers to maintain initial utility in the face of a tax-induced price change. The amount of compensation, which corresponds to the Hicksian measure of the *compensating variation* of the price change, may be calculated using the expenditure function as

$$E(\boldsymbol{p}_1, U_0) - E(\boldsymbol{p}_0, U_0) = \int_{\boldsymbol{p}_0}^{\boldsymbol{p}_1} \frac{\mathrm{d}E(\boldsymbol{p}, U_0)}{\mathrm{d}\boldsymbol{p}} \mathrm{d}\boldsymbol{p} = \int_{\boldsymbol{p}_0}^{\boldsymbol{p}_1} x^c(\boldsymbol{p}, U_0)\, \mathrm{d}\boldsymbol{p}, \tag{2.1}$$

which is well-defined even for a vector of changing prices $\boldsymbol{p}$ – the Hicksian variations are single-valued, regardless of the order of integration of the different price changes in Equation (2.1). For each market, this measure equals the area between prices p_0 and p_1 to the left of the compensated demand curve $D^c(U_0)$. Thus, the deadweight loss equals

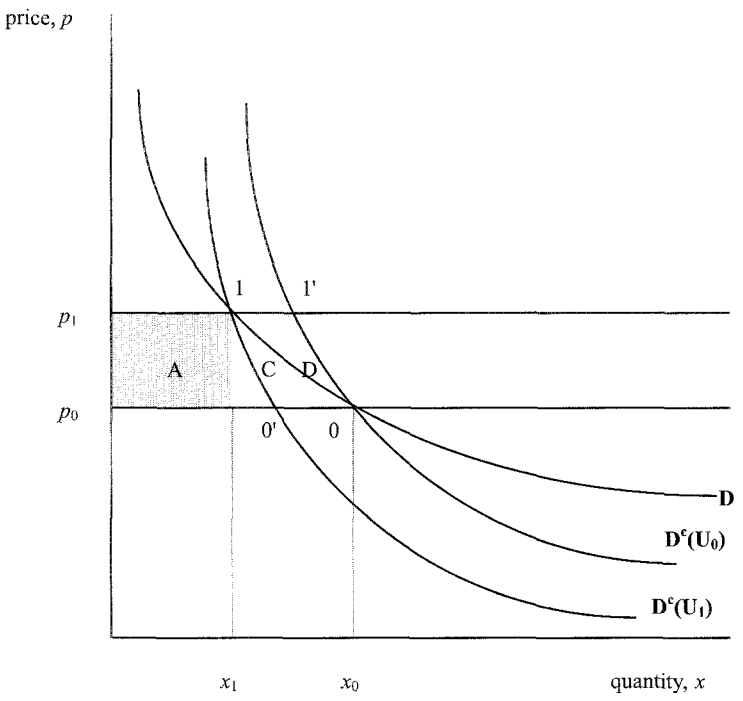


Fig. 2. Using Hicksian variations to measure excess burden.

area D in the figure – still approximately a "Harberger triangle", but different than that defined by the ordinary demand curve in Figure 1 [2].

An alternative conceptual experiment is to begin with the tax already in place and then remove it, extracting from consumers in lump-sum fashion an amount that prevents them from changing their utility levels while the tax is removed. Because the initial tax is distortionary, it is necessary to extract more from consumers than the tax revenue, the difference representing the excess burden of the initial tax. Starting from point 1 in Figure 2, this experiment follows the compensated demand curve $D^c(U_1)$ down to point 0′, where the price reaches its no-tax level but utility remains unchanged. Again using the expenditure function to calculate the amount the government extracts in this case – the Hicksian *equivalent variation,* based on the formula in Equation (2.1) with U_1 in place of U_0 – the amount equals the area to the left of demand curve $D^c(U_1)$ between prices p_0 and p_1. This exceeds the forgone

[2] Note that this definition is equally well-defined for the case of negative revenue, in which we would trace a path down the compensated demand curve from point 0. There, too, the tax system generates excess burden, in that the revenue lost exceeds the absolute value of the associated compensating variation. This serves as an important reminder that deadweight loss is the result of distortion, not of raising revenue *per se.*

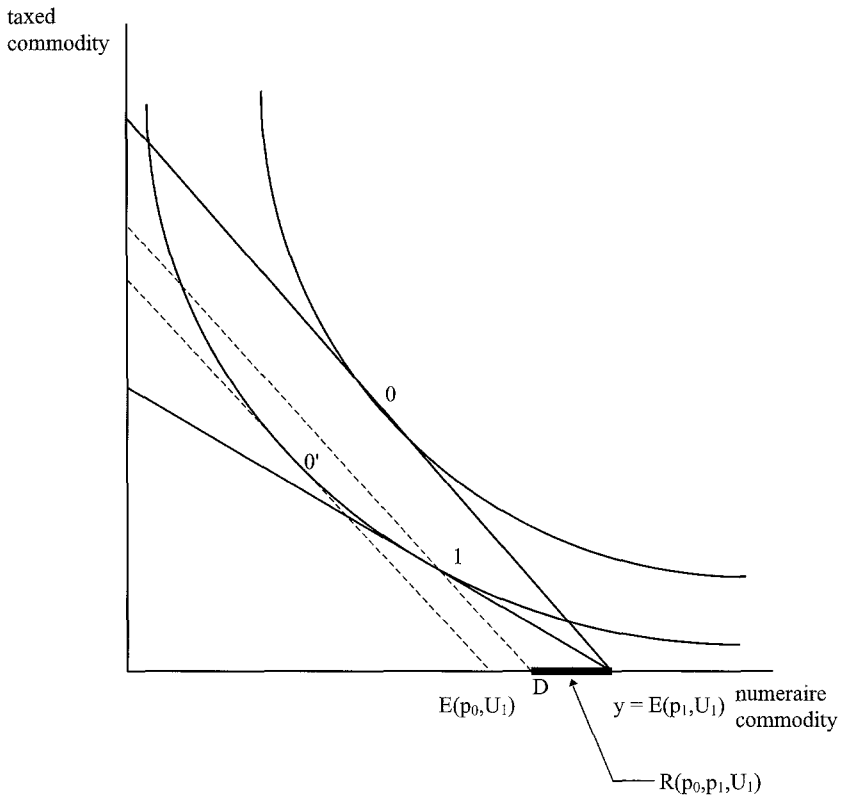


Fig. 3. Excess burden: an alternative graphical representation.

revenue – in this case the actual revenue defined by area A – and again does so by a "triangle".

Although these two measures are the most intuitive, they are actually just examples drawn from a class of measures based on arbitrary levels of utility, say U_i:

$$E(\boldsymbol{p}_1, U_i) - E(\boldsymbol{p}_0, U_i) - R(\boldsymbol{p}_0, \boldsymbol{p}_1, U_i), \tag{2.2}$$

where $R(\boldsymbol{p}_0, \boldsymbol{p}_1, U_i) \equiv (\boldsymbol{p}_1 - \boldsymbol{p}_0) \cdot x^c(\boldsymbol{p}_1, U_i)$ is the level of revenue collected with taxes in place and utility fixed at level U_i.

As Figure 3 shows, it is also possible to represent excess burden in a graph in commodity space. In the figure, the consumer's indifference curve is tangent to the original budget line at point 0, which corresponds to point 0 in Figure 2. The tax rotates the consumer budget line as shown, leading to consumption at point 1 (corresponding to point 1 in Figure 2), at which tax revenue, measured in terms of the numeraire commodity, equals $R(p_0, p_1, U_1)$. The consumer could maintain utility level U_1 in the absence of taxes by consuming at point 0′ (again, as labeled in Figure 2), where

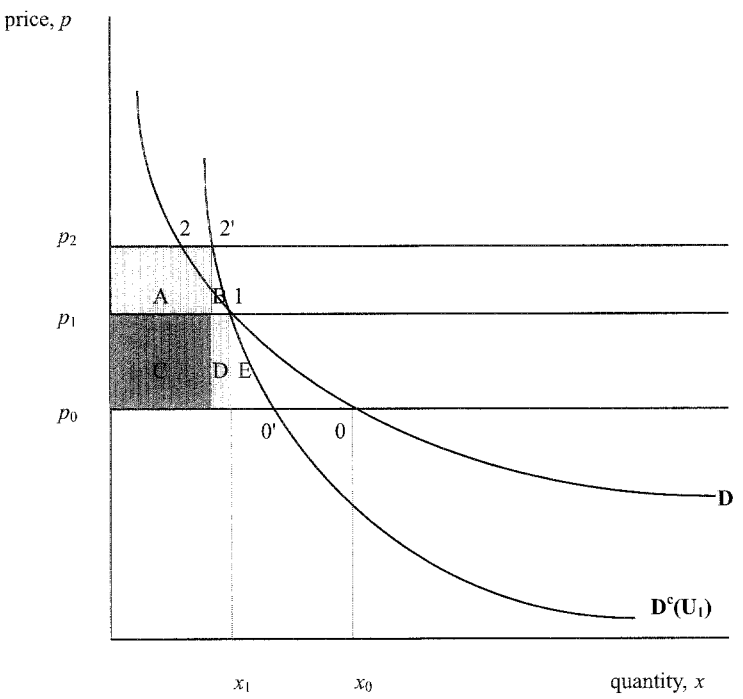


Fig. 4. Marginal excess burden of a pre-existing tax.

only $E(p_0, U_1)$ of expenditure would be required, which is less (as measured by the numeraire commodity) than the expenditure necessary to generate utility level U_1 when consumption is distorted by taxes (as it is at point D). The difference is the equivalent variation measure of excess burden, based on expression (2.2) for utility level U_1.

It is straightforward to generalize this class of measures to situations in which initial equilibria are not Pareto-optimal due to pre-existing taxes. The marginal excess burden of a tax change is the difference between the Hicksian variation associated with the price change and the *change* in tax revenue (which, in the absence of pre-existing taxes, is simply tax revenue), at the chosen level of utility:

$$E(\boldsymbol{p}_2, U_i) - E(\boldsymbol{p}_1, U_i) - [R(\boldsymbol{p}_0, \boldsymbol{p}_2, U_i) - R(\boldsymbol{p}_0, \boldsymbol{p}_1, U_i)], \tag{2.3}$$

in which $\boldsymbol{p}_2$ is the price vector after the tax change. For a given reference utility level U_i, this definition has the important property that the marginal excess burden in moving from point 1 to point 2 equals the difference between the excess burden at point 2 and the excess burden at point 1, as defined in expression (2.2).

Figure 4 illustrates this measure for the case in which an initial tax in a single market that changed the consumer price from p_0 to p_1 is then increased, raising the price to p_2. The figure illustrates the marginal excess burden of this tax increase, taking

the reference utility level to be that obtained at point 1, the consumption point with the initial tax in place. The Hicksian variation of the additional price change equals the sum of areas A and B. The change in tax revenue (with utility held constant) equals the difference between final tax revenue (areas A + C) and tax revenue prior to the imposition of the second tax, (C + D), or a difference of A − D. That is, with a pre-existing tax, it is necessary to net the revenue lost on forgone purchases against the revenue gained from a higher tax on remaining purchases. Thus, the marginal excess burden consists not only of the "triangle" B, but also the rectangle D. Marginal excess burden is no longer just a second-order phenomenon (the triangle) that vanishes with a small tax increase, but instead is of first-order significance. The total excess burden (calculated at utility level U_1) of both taxes equals this marginal excess burden plus the excess burden of the initial tax, equal to area E.

2.2. Variations in producer prices

The analysis thus far adopts the simplifying assumption of fixed relative producer prices, but it is possible to extend the various measures of excess burden to the more general case in which producer prices vary. It is helpful to begin with a graphical exposition. Figure 5 repeats the experiment of Figure 3, but does so in a case in which the relative producer price of the taxed good – the inverse slope of the production possibilities frontier (PPF), shown in bold – varies with the output mix.

Starting again at an equilibrium in which a distortionary tax is used to raise revenue from the representative household, the household's consumption bundle is shown at point 1, which corresponds to point 1 in Figure 3. Production occurs at point 1^P in the figure, and the government raises revenue in the numeraire commodity equal to the horizontal distance between points 1 and 1^P. The consumer price p_1 exceeds the producer price q_1 by the tax per unit of output. The household's income (in units of the numeraire commodity) is y_1, and its indifference curve is tangent to the consumer price line at point 1. Also passing through point 1 (but having a slope $-1/q_1$ and not tangent to the indifference curve) is a "private" production possibilities frontier – the original PPF, displaced to the left by the amount of the numeraire commodity corresponding to government consumption. Because the government is assumed to absorb only the numeraire commodity, this displacement is horizontal; otherwise, point 1 would not lie directly to the left of point 1^P. If, instead, the government devoted all tax revenues to purchases of the taxed commodity, then point 1 would lie directly *below* point 1^P. It should be clear that (unlike in the experiment with fixed producer prices) the equilibrium is affected by how the government uses its revenue, since government purchases influence relative demand and hence relative producer prices of the two commodities.

Excess burden is the amount of additional revenue the government could collect without harming the consumer, were lump-sum taxes used instead of distortionary taxes. It is necessary to specify the form that this extra revenue takes. Here, all revenue takes the form of the numeraire commodity, shifting the "private" PPF horizontally

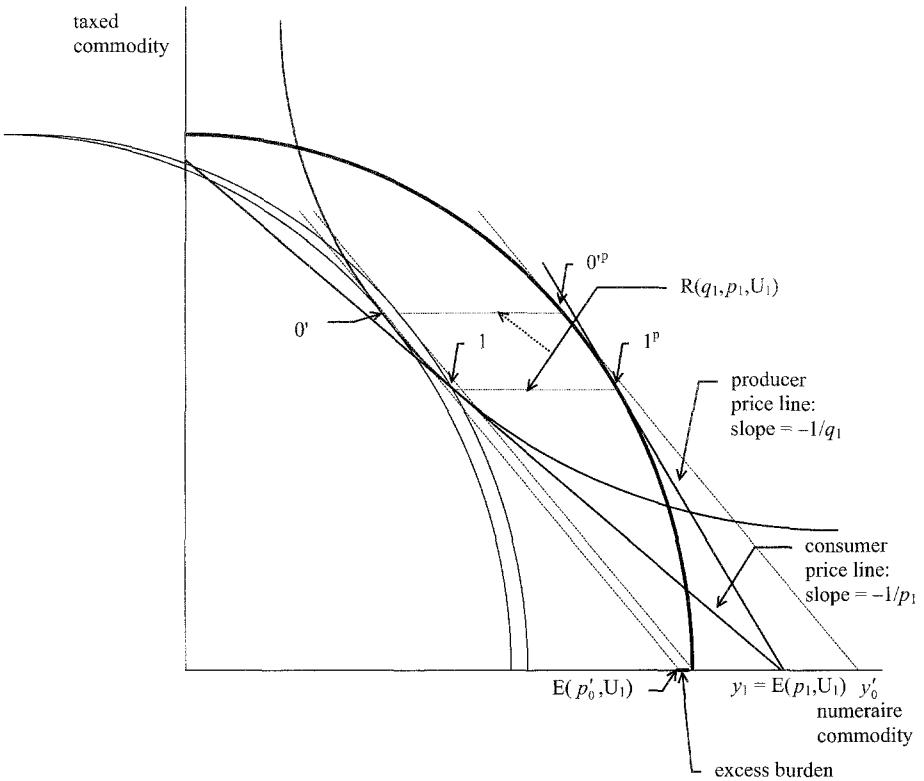


Fig. 5. Excess burden with varying producer prices.

to the left until tangent (at point 0′) with the indifference curve passing through point 1. Corresponding to consumption point 0′ is the production point $0'^{p}$. Excess burden is measured as the horizontal distance between this undistorted point 0′ and the corresponding point on the "private" PPF passing through point 1. Excess burden can be defined algebraically by noting that the horizontal distance between points 0′ and $0'^{p}$ equals the sum of excess burden and tax revenue (the same revenue as that raised in the initial equilibrium, $R(\boldsymbol{q}_1, \boldsymbol{p}_1, U_1)$. Thus, letting y'_0 be the value of the household's income from production at point 0′, excess burden equals

$$y'_0 - E(\boldsymbol{p}'_0, U_1) - R(\boldsymbol{q}_1, \boldsymbol{p}_1, U_1) = E(\boldsymbol{p}_1, U_1) - E(\boldsymbol{p}'_0, U_1) + y'_0 - y_1 - R(\boldsymbol{q}_1, \boldsymbol{p}_1, U_1), \tag{2.4}$$

with the last step in Equation (2.4) following from the identity that $E(\boldsymbol{p}_1, U_1) = E(\boldsymbol{p}_1, U(\boldsymbol{p}_1, y_1)) \equiv y_1$. As in the case with fixed producer prices, the measure defined

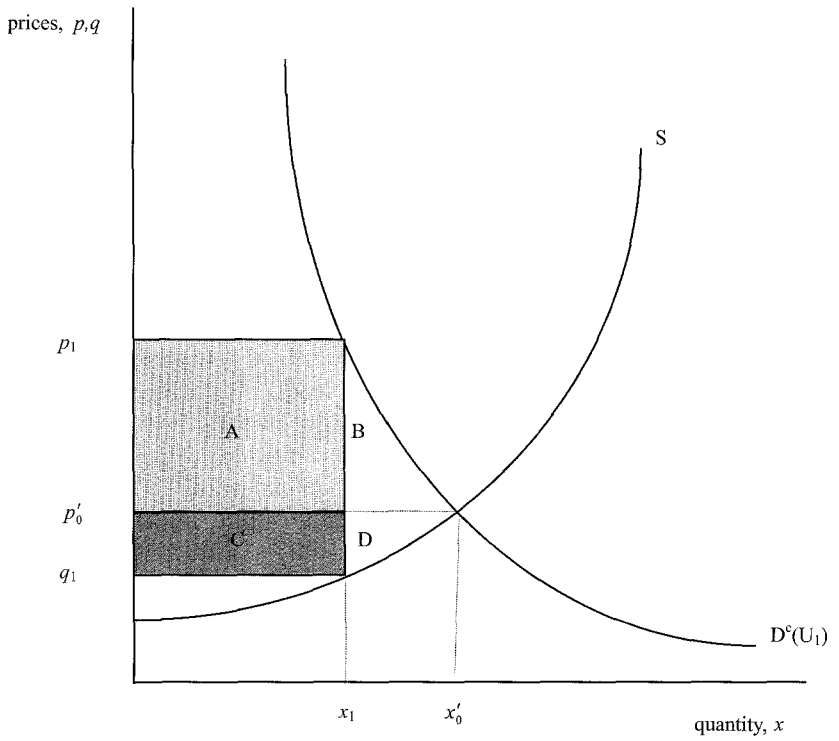


Fig. 6. Excess burden with an upward sloping supply curve.

in Equation (2.4) may be constructed for different reference utility levels[3]. Also, differences in excess burden as measured by Equation (2.4) correspond to changes in excess burdens due to additional taxes.

Expression (2.4) collapses to (2.2) when producer prices do not change, for then income y is fixed and the net-of-tax price vector in the tax-distorted equilibrium, $\boldsymbol{q}_1$, and the price vector in the undistorted equilibrium, $\boldsymbol{p}_0'$, both are identical to the original price vector $\boldsymbol{p}_0$. The extra term, $y_0' - y_1$, is the change in income along the production possibilities frontier when moving from point 1^p to point $0'^p$. By the envelope theorem, the change in income equals $\int_{\boldsymbol{q}_1}^{\boldsymbol{p}_0'} \boldsymbol{x}(\boldsymbol{q})\, d\boldsymbol{q}$, where $\boldsymbol{x}(\boldsymbol{q})$ is the quantity vector of goods produced at price vector $\boldsymbol{q}$. It is then possible to represent excess burden in a single market in price–quantity space, as does the diagram in Figure 6, in this case with an upward-sloping supply curve for the taxed good, $x(q)$. The excess burden, according to expression (2.4), equals the sum of Hicksian consumers' surplus, areas A + B, plus

[3] The expression for excess burden, and its graphical interpretation, becomes somewhat more complicated if the government absorbs both taxed and untaxed commodities. See Auerbach (1985) for further discussion.

the change in income, areas C+D (sometimes known as "producers' surplus") minus tax revenue, A+C, for a net excess burden of areas B+D.

For future reference, it is useful to present a very simple expression for the marginal excess burden of taxation. Totally differentiating the right-hand side of Equation (2.4) yields

$$\begin{aligned} dEB &= \frac{dE}{dp}dp_1 - \frac{dy}{dq}dq_1 - (p_1 - q_1)'\frac{dx^c}{dp}dp - x'(dp_1 - dq_1) \\ &= x^c(p_1, U_1)dp_1 - x(q_1)\,dq_1 - (p_1 - q_1)'\frac{dx^c}{dp}dp_1 - x'(dp_1 - dq_1) = -t'\frac{dx^c}{dp}dp_1, \end{aligned} \tag{2.5}$$

where the last step follows from the fact that $x^c(p_1, U_1) = x(q_1)$. That is, the change in excess burden equals the sum of the products of existing tax rates and changes in output. This result is extremely useful in searching for taxes that impose minimal excess burden. It is sometimes expressed as a first-order Taylor approximation for discrete changes, $-t'\Delta x$, or a second-order approximation $-(t'\Delta x + \frac{1}{2}\Delta t'\Delta x)$. The second-order approximation taken around the undistorted point ($t' = 0$), with Δt set equal to the tax vector itself, approximates a measure of the total excess burden of the tax system [e.g., Harberger (1964a)]. From this approximation comes the common intuition that excess burden increases with the square of a tax. If one considers the second-order approximation for a single tax Δt_i and producer prices fixed, excess burden is $-\frac{1}{2}\Delta t_i(dx_i^c/dt_i)\Delta t_i$.

2.3. *Empirical issues in the measurement of excess burden*

While the theory of deadweight-loss measurement has a long and colorful history that dates back to the nineteenth-century contributions of Jules Dupuit (1844) and Fleeming Jenkin (1871/72), economists seldom measured actual deadweight losses prior to the pioneering work of Arnold Harberger in the 1950s and 1960s. In two influential papers published in 1964, Harberger (1964a) derived the approximation (2.5) used to measure deadweight loss and (1964b) applied the method to estimate deadweight losses due to income taxes in the United States. Harberger shortly thereafter (1966) produced estimates of the welfare cost of US capital taxes. A generation of empirical studies by other scholars followed the publication of Harberger's subsequent survey article (1971)[4].

The empirical work that followed Harberger's efforts focussed on the use of simple deadweight-loss formulas to estimate the welfare impact of a wide array of tax-induced distortions, including those to labor supply [Browning (1975), Hausman (1981a)], saving [Feldstein (1978)], corporate taxation [Shoven (1976)], and the consumption of goods, such as housing and non-housing consumption items, that are taxed to differing degrees [King (1983)][5]. In addition, some attention was devoted to refining

[4] See Hines (1999) for an interpretive survey of this literature.

[5] See the discussion in Auerbach (1985) and the more recent survey by Slesnick (1998).

the approximations used in applying estimated behavioral parameters to calculate deadweight losses. The variant of Equation (2.5) used by Harberger, in which a form of uncompensated demand is used in place of compensated demand, approximates a compensated measure of welfare change (2.4). One question of interest to subsequent investigators is the practical difference between results obtained using Harberger-style approximations and those available from more exact measures. As Mohring (1971) and subsequent authors note, it is often the case that the same demand information necessary to calculate approximations to Equation (2.5) can, if properly modified, be used to calculate Hicksian deadweight-loss measures of the form (2.4). The extent to which these two methods generate different answers is, of course, an empirical question. Rosen (1978) finds that Equation (2.4) and approximations to (2.5) track each other rather closely, but Hausman (1981b) offers some examples in which they differ considerably.

The generation of empirical work following Harberger calls attention to the importance of linking the strategy used to estimate demand and the ultimate goal of using the estimates to perform welfare analysis. Specifically, this entails estimating models that can be integrated to obtain expenditure functions from which expressions such as Equation (2.4) can be derived[6]. In the course of performing such estimation, it is of course desirable to make the model sufficiently flexible that its functional form imposes as few answers as possible. For this purpose it can be useful to employ algorithms that estimate expenditure functions numerically based on demand-parameter estimates [Vartia (1983)].

A major practical difficulty in measuring the excess burden of a single tax, or of a system of taxes, is that excess burden is a function of demand interactions that are potentially very difficult to measure. For example, a tax on labor income is expected to affect hours worked, but may also affect the accumulation of human capital, the intensity with which people work, the timing of retirement, and the extent to which compensation takes tax-favored (e.g., pensions, health insurance, and workplace amenities) in place of tax-disfavored (e.g., wage) form. In order to estimate the excess burden of a labor-income tax, it is in principle necessary to estimate the effect of the tax on these and other decision margins. Analogous complications are associated with estimating the excess burdens of most other taxes. In practice, it can be very difficult to obtain reliable estimates of the impact of taxation on just one of these variables.

It is in reaction to the complicated nature of the problem of separately estimating the effect of taxation on all of a taxpayer's decision margins that a number of recent papers estimate variants of Equation (2.5) in which the dependent variable is taxable

[6] Examples of such estimation strategies include Deaton and Muellbauer (1980), Gallant (1981), and Jorgenson, Lau and Stoker (1982). Hausman and Newey (1995) offer a nonparametric alternative.

income. The usefulness of this formulation is evident from considering the consumer's problem in maximizing

$$U(x_1, x_2, x_3, l), \tag{2.6}$$

in which x_1, x_2 and x_3 are commodities taxed to differing degrees, and l is leisure. In order to illustrate the issues involved, we consider the case in which good 1 is an ordinary commodity that consumers purchase out of after-tax income, purchases of good 2 are fully deducted from taxable income, and purchases of good 3 are partially deductible for tax purposes. Given a labor endowment of $\tilde{L}$, a wage of w, and facing a (flat-rate, for purposes of simplicity) labor-income tax rate of τ, the consumer's budget constraint is

$$p_1 x_1 + p_2 x_2 (1-\tau) + p_3 x_3 (1-\alpha\tau) + w(1-\tau) l \leqslant w(1-\tau)\tilde{L}, \tag{2.7}$$

in which α denotes the degree to which purchases of x_3 are deductible for tax purposes. Feldstein (1999) notes that the budget constraint (here, 2.7) can be transformed to yield

$$\frac{p_1 x_1 + (1-\alpha) p_3 x_3}{1-\tau} \leqslant w(\tilde{L} - l) - p_2 x_2 - \alpha p_3 x_3. \tag{2.8}$$

The right-hand side of constraint (2.8) equals taxable income, since labor effort is given by $(\tilde{L} - l)$, purchases of commodity 2 are deductible from income, and a fraction α of purchases of commodity 3 is also deductible. In this environment, higher labor-income tax rates create deadweight loss by discouraging consumption of good 1, and partially discouraging consumption of good 3, relative to consumption of leisure and of good 2. It is therefore possible to estimate deadweight loss by estimating the responsiveness of taxable income to changes in tax rates, since doing so traces the effect of changes in τ on the numerator of the left-hand side of constraint (2.8).

Several empirical studies, including Lindsey (1987), Feldstein (1995), Auten and Carroll (1999), Goolsbee (2000) and Moffitt and Wilhelm (2000), consider the responsiveness of taxable income to tax rates, relying on major US tax changes to provide variation in tax rates. The American tax reforms of 1981 and 1986 significantly reduced marginal tax rates, particularly those of high-income taxpayers, while tax reforms enacted in 1990 and 1993 had the opposite effect of raising tax rates on high-income taxpayers. The evidence indicates that taxable income is generally very responsive to tax changes, with estimated response elasticities that significantly exceed the typically very modest estimated effects of taxation on numbers of hours worked. Lindsey and Feldstein report elasticities of taxable income in excess of unity, while Auten and Carroll, Goolsbee, and Moffitt and Wilhelm provide a range of somewhat more modest estimates. All of these studies report that the taxable incomes of high-income taxpayers are far more responsive to tax-rate changes than are the taxable incomes of the rest of the population.

There are two important considerations in interpreting this evidence. The first is that, in order to use the framework described by constraint (2.7) as the basis of analysis, it is important to estimate the responsiveness to taxation of the present value of taxable income. Tax avoidance often takes the form of deferring a tax obligation from one period into another in order to reduce its present value. Consequently, the reaction of short-term taxable income to a tax change may exceed the reaction of the present value of taxable income, which Goolsbee (2000) finds occurred with executive compensation in response to the 1993 US tax change. In addition to the difficulty of distinguishing empirically short-term from long-term reactions, there is the added complication that timing behavior depends on anticipated future tax policies that may not be known to the analyst.

The second consideration is that tax changes that reduce one type of taxable income may have offsetting or reinforcing effects on other sources of taxable income. For example, increasing the personal income tax rate may encourage some high-income taxpayers to incorporate their personal businesses, thereby reducing total income earned by individuals through proprietorships while increasing corporate income. A simple calculation of the responsiveness of personal income to changes in personal income tax rates would then overstate the true effect of tax changes on total taxable income. Furthermore, individuals purchase commodities that are taxed to differing degrees, and tax collections from these sources are appropriately included in reactions to tax changes[7]. Properly accounting for all of these reactions when performing welfare analysis is a daunting task, but one that is more likely than many of the available alternatives to provide useful answers.

3. The design of optimal taxes

Taxes (other than lump-sum taxes) distort behavior, yet society needs to collect revenue to pursue various social objectives. The optimal-taxation literature identifies tax systems that minimize the excess burden of taxation, subject to various restrictions on tax instruments and information available to the government, and under different assumptions about population heterogeneity and the functioning of private markets.

Historically, there are three strands in the development of the optimal-taxation literature. One, initiated by the seminal work of Ramsey (1927) and carried on, perhaps most notably, by Diamond and Mirrlees (1971a,b), concentrates on the design of commodity taxes. A second set of contributions, beginning with Mirrlees (1971), considers more general nonlinear income taxes and focuses on the role of such taxes in addressing distributional concerns. Finally, the work of Pigou (1947) and others

[7] Note that Equation (2.7) would be unchanged if expenditures on commodity 3 were nondeductible, but purchases of commodity 3 were subject to an ad valorem tax at rate $(-\alpha\tau)$. As a general matter, however, pre-existing distortions due to taxes, imperfect competition, and other sources of divergence between price and marginal cost should be incorporated in measuring deadweight loss.

analyzes the use of taxes to address two types of market failures: financing "public" goods not provided by the private sector, and correcting externalities associated with incomplete private-sector markets[8]. Although these three strands in the literature have converged, it is still useful to consider them separately in turn before discussing their interrelationship.

3.1. The Ramsey tax problem

The simplest version of the Ramsey tax problem abstracts from population heterogeneity and posits that the government must raise a fixed sum of tax revenue with proportional commodity taxes, leaving to the side how such revenue is to be spent. With a population of identical individuals, typically analyzed as a single representative individual, the goal of optimal tax design is to minimize the excess burden associated with raising the needed revenue. We typically rationalize government's inability to use lump-sum taxes by saying that such taxes are inequitable, although this may seem a bit forced in a setting with identical individuals. It may help to think of this simple problem as a necessary building block, rather than as one that adequately models a realistic situation.

The representative consumer maximizes utility, $U(\boldsymbol{x})$, over a vector of commodities x_i $(i=0,1,\ldots,N)$, subject to the budget constraint $\boldsymbol{p}\cdot\boldsymbol{x} \leqslant y$, where $\boldsymbol{p}$ is the corresponding vector of consumer prices and y is lump-sum income. To raise the required level of revenue, R, the government imposes a vector of taxes on the commodities, $\boldsymbol{t}$, driving a wedge between consumer prices and producer prices, $\boldsymbol{q}$. It is useful to assume initially that this vector of producer prices is fixed (perhaps by world prices), but as will be seen later, this is not a restrictive assumption in characterizing the optimum. With given producer prices, the government in setting tax rates is effectively choosing the consumer price vector, since $\boldsymbol{p}=\boldsymbol{q}+\boldsymbol{t}$. Thus, the government's optimal tax problem can be modeled as

$$\max_{p} V(\boldsymbol{p},y) \quad \text{subject to} \quad (\boldsymbol{p}-\boldsymbol{q})'\boldsymbol{x} \geqslant R, \tag{3.1}$$

where $V(\cdot)$ is the household's indirect utility function.

[8] One potentially important market failure not considered by this chapter is the incompleteness of markets in state-contingent claims that might otherwise be used to diversify risks. In such a setting, it is possible for taxation to improve welfare simply by reducing (after-tax) private returns – since the government can pool risks through its tax and spending actions. Diamond, Helms and Mirrlees (1980), Varian (1980) and Eaton and Rosen (1980) analyze the properties of optimal distortionary taxation in stochastic settings with missing state-contingent markets, while Sandmo (1985) provides a more general survey of the impact of taxation in settings characterized by risk.

To see the relationship between the optimal tax problem and the problem of excess burden, note that the problem (3.1) is equivalent to

$$\min_{\boldsymbol{p}} y - E(\boldsymbol{q}, V(\boldsymbol{p},y)) - R \quad \text{subject to} \quad (\boldsymbol{p}-\boldsymbol{q})'\boldsymbol{x} \geqslant R, \tag{3.2}$$

because y and R are constants and $E(\boldsymbol{q}, V(\boldsymbol{p},y))$ is monotonically increasing in $V(\boldsymbol{p},y)$. But, as $y \equiv E(\boldsymbol{p}, V(\boldsymbol{p},y))$, expression (3.2) amounts to minimizing the excess burden of taxation subject to the revenue constraint, in which excess burden is evaluated at the utility level $V(\boldsymbol{p},y)$ that holds in the presence of taxation (that based on the Hicksian equivalent variation)[9].

Without further restrictions, the optimal tax problem is actually quite trivial, since excess burden can be avoided entirely simply by raising all prices by a uniform multiple. That is, let $\boldsymbol{p} = \phi\boldsymbol{q}$, with $\phi > 1$ chosen so that $(\phi - 1)\boldsymbol{q}'\boldsymbol{x} = R$. Then excess burden is

$$\begin{aligned} &E(\phi\boldsymbol{q}, V(\phi\boldsymbol{q},y)) - E(\boldsymbol{q}, V(\phi\boldsymbol{q},y)) - (\phi-1)\,\boldsymbol{q}'\boldsymbol{x}(\phi\boldsymbol{q},y) \\ &\quad = \phi E(\boldsymbol{q}, V(\phi\boldsymbol{q},y)) - E(\boldsymbol{q}, V(\phi\boldsymbol{q},y)) - (\phi-1)\,\boldsymbol{q}'\boldsymbol{x}(\phi\boldsymbol{q},y) \\ &\quad = (\phi-1)\,\boldsymbol{q}'\boldsymbol{x}(\phi\boldsymbol{q},y) - (\phi-1)\,\boldsymbol{q}'\boldsymbol{x}(\phi\boldsymbol{q},y) = 0, \end{aligned} \tag{3.3}$$

where the second step follows from the fact that the expenditure function is homogeneous of degree 1 with respect to prices, and the third step from the identity $E(\boldsymbol{q}, V(\phi\boldsymbol{q},y)) \equiv \boldsymbol{q}'\boldsymbol{x}^{c}(\phi\boldsymbol{q}, V(\phi\boldsymbol{q},y)) = \boldsymbol{q}'\boldsymbol{x}(\phi\boldsymbol{q},y)$.

Raising revenue in this way entails no excess burden because it is equivalent to imposing a lump-sum tax; the household's budget constraint in the presence of uniform taxation is

$$\phi\boldsymbol{q}'\boldsymbol{x} = y \quad \Rightarrow \quad \boldsymbol{q}'\boldsymbol{x} = y - (\phi-1)\,y/\phi. \tag{3.4}$$

Thus, it is necessary to impose taxes that create excess burden only if it is impossible to adjust the tax rates freely on all $N+1$ commodities, or else if exogenous income $y=0$, in which case uniform taxes raise no revenue[10].

What does it mean for consumers to have no exogenous income? The interpretation of the condition that $y=0$ depends on the definition of commodities $\boldsymbol{x}$. Consider, for example, the simple case of three commodities, including two that the household purchases, x_1 and x_2, and a third, labor, that the household supplies as a factor to the production process. It is customary to write the budget constraint for this problem as

$$p_1x_1 + p_2x_2 + wl = w\tilde{L}, \tag{3.5}$$

where l is leisure consumed and $\tilde{L}$ is the household's time endowment. Households divide their time between leisure and working at a wage of w per unit of working time.

[9] This measure of excess burden based on the equivalent variation may be used more generally to compare any two tax systems, neither of which is necessarily optimal. This property has led some [e.g., Kay (1980)] to prefer its use over measures based on other reference-utility levels.

[10] Note that if $y < 0$, it is possible to raise revenue with uniform taxation by choosing $\phi < 1$.

With the budget constraint written this way, it is clear that a uniform tax on consumption and leisure is equivalent to a lump-sum tax on the household's time endowment. It is standard to rule this out by specifying that leisure cannot be taxed, that the government is restricted to taxing labor, $L = \tilde{L} - l$. With such a restriction, if leisure is taxed, the government must offer a matching subsidy to the time endowment, a requirement that eliminates the possibility of lump-sum taxation. That is, Equation (3.5) can be rewritten as

$$p_1 x_1 + p_2 x_2 + w(l - \tilde{L}) = p_1 x_1 + p_2 x_2 - wL = 0, \tag{3.6}$$

in which it is clear that uniform taxes on x_1, x_2 and L raise no revenue. This result may seem counterintuitive because the "tax" on the household's leisure purchases *raises* the price of labor, corresponding to what we normally think of as a wage *subsidy*. It is possible to raise revenue by lowering the wage while raising prices p_1 and p_2, but this no longer leaves relative prices undistorted – it lowers the real wage in terms of each consumption good. Indeed, a labor-income tax and a uniform tax on the two consumption goods are equivalent tax policies. With the budget constraint expressed as

$$p_1 x_1 + p_2 x_2 = wL, \tag{3.7}$$

it is clear that raising commodity prices is the same policy as reducing wages.

Thus, the need to use distortionary taxes results either from a restriction on the use of tax instruments (e.g., it is not possible to tax leisure, or the consumption of any other endowed commodity, separately from its endowment) or on the absence of exogenous income (if labor, rather than leisure, is the relevant commodity). Because it is standard to assume that the government cannot impose separate taxes on endowments in labor or other commodities[11], it is easier to adopt the second interpretation, expressing commodities as flows between the household and production sectors and leaving only "pure" economic rent potentially on the right-hand side of the budget constraint.

With no lump-sum income, two tax systems are equivalent if they differ by proportional taxes on all commodities. Without lump-sum income one is therefore free to normalize one of the taxes, say on good 0, to zero, and for convenience choose the

[11] It is customary simply to assume that the government cannot tax an individual's labor endowment because this endowment is not observable; equivalently, we assume that we can observe an individual's labor income, but not the effort expended or leisure forgone in earning that income. Although there has been some work considering modifications of this assumption [e.g., Stern (1982)], this issue has received relatively little attention in the literature.

same good as numeraire, i.e., $q_0 = p_0 = 1$. The maximization problem (3.1), with the multiplier μ associated with the budget constraint, yields N first-order conditions:

$$-\lambda x_i + \mu \left[x_i + \sum_j t_j \frac{dx_j}{dp_i} \right] = 0, \qquad i = 1, \ldots, N, \tag{3.8}$$

in which $\lambda \equiv \partial V(\boldsymbol{p}, y) / \partial y$ is the marginal utility of income. Making use of the Slutsky decomposition, Equation (3.8) implies

$$\sum_j t_j S_{ji} = -\frac{(\mu - \alpha)}{\mu} x_i, \qquad i = 1, \ldots, N, \tag{3.9}$$

where S_{ji} is the jith element of the Slutsky matrix $S \equiv d\boldsymbol{x}^c / d\boldsymbol{p}$, and $\alpha = \lambda + \mu \sum_j t_j \frac{dx_j}{dy}$ is the "social" marginal utility of income that includes the value of the additional tax revenue raised when the household receives another unit of income[12].

Although there is no independent condition for good 0, it may be shown [see Auerbach (1985)] that the N first-order conditions in Equation (3.9) imply a comparable condition for good 0, a result that should not be too surprising given that the choice of the good to bear the zero tax is arbitrary. Stacking these $N + 1$ conditions yields

$$S\boldsymbol{t} = -\left(\frac{\mu - \alpha}{\mu} \right) \boldsymbol{x}. \tag{3.10}$$

Premultiplying both sides of Equation (3.10) by the tax vector $\boldsymbol{t}'$, we obtain an equation in which the left-hand side is a negative semi-definite quadratic form and the right-hand side equals the product of the constant term $-(\mu - \alpha)/\mu$ and tax revenue $\boldsymbol{t}'\boldsymbol{x}$ [13]. Thus, if revenue is positive, $\mu \geqslant \alpha$ – the marginal social cost of raising additional revenue, μ, is at least as large as the cost of raising revenue in lump-sum fashion, α, i.e., marginal excess burden is nonnegative. This condition does not hold for arbitrary tax schedules, but starting from an optimal tax system for any given level of revenue means that there is no opportunity to reduce excess burden while raising taxes, for example by bringing up the tax rates on goods that initially are undertaxed[14]. Note

[12] Samuelson (1951) uses the symmetry of the Slutsky matrix ($S_{ij} = S_{ji}$) to interpret Equation (3.9) as implying that optimal taxes entail equiproportionate compensated reductions in demands for all commodities. While valid locally, this interpretation relies on constancy of the elements of the Slutsky matrix as tax rates change, a feature they do not generally exhibit.

[13] Because the first element of the tax vector is zero, the relevant part of the Slutsky matrix is the submatrix formed by striking the first row and column of S. This submatrix and the associated quadratic form will generally be negative definite, as long as some of the omitted substitution terms are nonzero.

[14] Note that marginal excess burden is nonpositive when revenue is initially negative, because raising revenue means reducing the level of distortions caused by subsidies.

that this inequality relates μ to α, not to λ, the private marginal utility of income. By the definition of α, $\mu \geqslant \alpha \Rightarrow \mu \geqslant \lambda$ only if revenue is nondecreasing in income, i.e., if the tax base is a normal composite good. This distinction is important to keep in mind when considering the literature that seeks to identify the "marginal cost of funds."

Before interpreting expression (3.10) further, it is useful to consider the more general case of variable producer prices.

3.2. *Changing producer prices*

Since the excess burden of a tax is a function of the extent to which the tax changes producer prices, it follows intuitively that allowing producer prices to vary alters the first-order conditions for the optimal tax schedule. Let the general production be characterized by

$$f(\boldsymbol{z}) \leqslant 0, \tag{3.11}$$

where $\boldsymbol{z}$ is the production vector, and perfect competition insures that $q_i/q_j = f_i/f_j \,\forall i, j$. Without loss of generality, the units of the production function can be chosen such that $q_i = f_i$. If there are constant returns to scale, then $f(\cdot)$ is homogeneous of degree zero in $\boldsymbol{z}$. Otherwise, there may be pure profits, $y = \boldsymbol{q}'\boldsymbol{z} > 0$.

With changing producer prices, it is not appropriate to specify the constraint in the optimal tax problem as a scalar value of tax revenue to be collected, so it is necessary to posit that the government absorbs a vector $\boldsymbol{R}$ of commodities. This implies that the consumption vector $\boldsymbol{x}$ satisfies $f(\boldsymbol{x} + \boldsymbol{R}) \leqslant 0$, thereby incorporating both revenue and production constraints. The optimal tax problem, then, is to maximize the indirect utility function $V(\boldsymbol{p}, y)$ subject to this constraint, and not that given in Equation (3.2). The associated Lagrangean expression is

$$V(\boldsymbol{p}, y) - \mu f(\boldsymbol{x} + \boldsymbol{R}), \tag{3.12}$$

and the government's problem is still that of choosing the consumer price vector $\boldsymbol{p}$, rather than the tax vector $\boldsymbol{t}$, even though the relationship between changes in the two vectors is more complicated than when producer prices are fixed[15]. The resulting first-order conditions are (recalling the normalization that $q_i = f_i$)

$$-\lambda x_i + \lambda \frac{\mathrm{d}y}{\mathrm{d}p_i} + \mu \left[-\sum_j q_j \frac{\mathrm{d}x_j}{\mathrm{d}p_i} \right] = 0, \qquad i = 1, \ldots, N. \tag{3.13}$$

[15] As discussed in Auerbach (1985), $\mathrm{d}\boldsymbol{p}/\mathrm{d}\boldsymbol{t} = [I - HS]^{-1}$, where H is the Hessian of $f(\cdot)$, so there is a one-to-one relationship between changes in $\boldsymbol{t}$ and changes in $\boldsymbol{p}$ as long as $[I - HS]$ is of full rank.

Differentiating the household's budget constraint $\boldsymbol{p}'\boldsymbol{x} = y$ with respect to p_i yields

$$x_i + \sum_j p_j \frac{\mathrm{d}x_j}{\mathrm{d}p_i} - \frac{\mathrm{d}y}{\mathrm{d}p_i} = 0, \qquad i = 1, \ldots, N, \tag{3.14}$$

and adding the left-hand side of this equation to the expression inside the brackets in Equation (3.13) yields

$$-\lambda x_i + \lambda \frac{\mathrm{d}y}{\mathrm{d}p_i} + \mu \left[x_i + \sum_j t_j \frac{\mathrm{d}x_j}{\mathrm{d}p_i} - \frac{\mathrm{d}y}{\mathrm{d}p_i} \right] = 0, \qquad i = 1, \ldots, N. \tag{3.15}$$

Since producer prices, and hence profits, change with $\boldsymbol{p}$, the derivative $\mathrm{d}x_j/\mathrm{d}p_i$ in Equation (3.15) includes the indirect effect of p_i on profits through changes in production:

$$\frac{\mathrm{d}x_j}{\mathrm{d}p_i} = \frac{\partial x_j}{\partial p_i} + \frac{\mathrm{d}x_j}{\mathrm{d}y} \cdot \frac{\mathrm{d}y}{\mathrm{d}p_i}. \tag{3.16}$$

Using this and the Slutsky decomposition, Equation (3.15) can be rewritten as

$$-\sum_j t_j S_{ji} = \frac{(\mu - \alpha)}{\mu} \left(x_i - \frac{\mathrm{d}y}{\mathrm{d}p_i} \right), \qquad i = 1, \ldots, N, \tag{3.17}$$

which differs from expression (3.9), the first-order condition in the case of fixed producer prices, by the term $\mathrm{d}y/\mathrm{d}p_i$ on the right-hand side. Thus, if there are constant returns to scale ($y \equiv 0$), the first-order conditions are identical [Diamond and Mirrlees (1971a,b)]. The same is true if the government imposes a pure profits tax, so that the after-tax value of y accruing to households is uniformly zero [Stiglitz and Dasgupta (1971)].

From expression (2.5), the left-hand side of Equation (3.17) equals the marginal excess burden associated with an increase in p_i. The second term on the right-hand side of Equation (3.17) is the net compensation required to maintain the individual's utility as p_i rises[16] which, by definition, exceeds the marginal revenue raised by the marginal excess burden induced by the price change. Thus, Equation (3.17) says that

[16] This term equals $-\frac{\mathrm{d}V(\boldsymbol{p},y)/\mathrm{d}p_i}{\mathrm{d}V(\boldsymbol{p},y)/\mathrm{d}y}$; according to Roy's identity, this equals the net increase in income required to maintain the household's utility level as p_i increases.

the excess burden of a marginal increase in any tax must be proportional to the sum of marginal revenue plus marginal excess burden, or:

$$\frac{dEB}{dp_i} = \frac{(\mu - \alpha)}{\mu}\left(\frac{dR}{dp_i} + \frac{dEB}{dp_i}\right), \qquad i = 1, \ldots, N. \tag{3.18}$$

It follows that the marginal excess burden per dollar of revenue raised, $(\mu - \alpha)/\alpha$, is also constant,

$$\frac{dEB}{dp_i} = \frac{(\mu - \alpha)}{\alpha}\frac{dR}{dp_i}, \qquad i = 1, \ldots, N, \tag{3.19}$$

which is an intuitive condition for minimizing the total excess burden induced by raising a given amount of revenue from alternative sources.

3.3. *The structure of optimal taxes*

The optimal tax rules just derived generally do not imply that the government should impose taxes at uniform rates, even in the simple case in which producer prices are fixed. For example, consider the three-good case, in which the two first-order conditions (3.9) yield

$$\frac{t_1}{t_2} = \frac{-S_{22}x_1 + S_{12}x_2}{-S_{11}x_2 + S_{21}x_1}, \tag{3.20}$$

which, using the fact that $\sum_j p_i S_{ij} = 0$, and defining $\theta_i \equiv t_i/p_i$ as the tax rate on good i, may be rewritten as

$$\frac{\theta_1}{\theta_2} = \frac{\varepsilon_{20} + \varepsilon_{21} + \varepsilon_{12}}{\varepsilon_{10} + \varepsilon_{21} + \varepsilon_{12}}, \tag{3.21}$$

where ε_{ij} is the compensated cross-price elasticity of demand for good i with respect to the price of good j.

This expression indicates that two goods should be taxed at equal rates (i.e., $\theta_1 = \theta_2$) if and only if the goods are equally complementary with respect to the untaxed good 0. The intuition sometimes offered for this result comes from the case in which the untaxed good 0 is labor, making it desirable to tax more heavily the good that is more complementary with leisure because it is impossible to tax leisure directly. But since expression (3.20) would also apply if a consumption good were chosen to bear the zero tax, it may be more accurate to say that complements to untaxed goods are taxed more heavily to achieve reductions in the untaxed goods without taxing them directly.

In the special case of zero cross-elasticities among all taxed goods, the first-order conditions (3.9) yield the "inverse elasticity rule" that $\theta_i \propto 1/\varepsilon_i$, since in this case each good's demand responds only to its own tax, so achieving a reduction of equal proportion means keeping $\theta_i \varepsilon_i$ constant.

3.4. An example

Suppose that household preferences over goods and leisure are described by the Stone–Geary utility function,

$$U(x_1, x_2, l) = (x_1 - a_1)^{\beta_1}(x_2 - a_2)^{\beta_2} l^{1-\beta_1-\beta_{12}}. \tag{3.22}$$

For this utility function, the cross elasticity ε_{i0} equals $(1-\beta_1-\beta_2)(1-a_i/x_i)$, so optimal taxes fall more heavily on the consumption good whose "basic need" a_i represents a larger portion of total consumption x_i. In terms of underlying preferences, it can be shown that this is equivalent to taxing more heavily the good with the higher value of $p_i a_i/\beta_i$, the good for which expenditures on basic needs are a greater fraction of the good's discretionary budget share, β_i. In the special case where $a_1 = a_2 = 0$, the Stone–Geary utility function collapses to the Cobb–Douglas function, and uniform taxes are optimal. The Cobb–Douglas utility function is separable into goods and leisure (or, to be more exact, into the taxed and untaxed commodities) and homogenous in goods – it can be written in the form $U(\phi(\boldsymbol{x}), l)$, where $\phi(\cdot)$ is a homogeneous function. This homothetic separability is a sufficient condition for uniform taxation [Atkinson and Stiglitz (1972)]. Separability alone does not suffice – as the general Stone–Geary example illustrates.

3.5. The production efficiency theorem

All of the tax instruments considered so far are proportional taxes on transactions between the household sector and the production sector. Production itself is assumed to face no distortions, and perfect competition ensures that the economy achieves a point on the production frontier. However, the government has access to policies that distort production while raising revenue, either through explicit taxes or through government production schemes that allocate inputs and outputs on the basis of criteria possibly different than those used by the private sector. One might think that such policy instruments would favorably augment the government's options, but this may well not be so.

Consider the case in which there is a second production sector, say controlled directly by the government, with production function $g(\cdot)$ and production vector $\boldsymbol{s}$, with the production set defined by $g(\boldsymbol{s}) \leqslant 0$. Distortions between the two sectors occur implicitly through the government's choice of the vector $\boldsymbol{s}$, with each sector, but not necessarily the two sectors in combination, assumed to be on its own production frontier. Further assume that production in both sectors is subject to constant returns to scale.

Because private production now equals the difference between purchases $\boldsymbol{x} + \boldsymbol{R}$ and government production $\boldsymbol{s}$, the government's problem is to maximize $V(\boldsymbol{p}, y)$ subject to $f(\boldsymbol{x} + \boldsymbol{R} - \boldsymbol{s}) \leqslant 0$ and $g(\boldsymbol{s}) \leqslant 0$. Forming the Lagrangean as before, with the multiplier ζ associated with the second sector's production, we obtain the same first-order conditions as before with respect to $\boldsymbol{p}$, and the conditions that $\mu f_i - \zeta g_i = 0 \; \forall i$

with respect to the vector s. This implies that all marginal rates of substitution in production should be equal, $f_i/f_j = g_i/g_j$, i.e., production should not be distorted. This result does not hold if pure profits are received by the household, and this helps provide insight into why it *does* hold when no such profits are received. In this special case, all household decisions are based on the relative price vector $\boldsymbol{p}$. It is possible to bring about any configuration of this vector that is consistent with the revenue constraint, without resorting to production distortions. Thus, production distortions can serve only to reproduce what can already be achieved, but with the additional social cost of lost production. Of course, if the government is not free to adjust all relative prices directly, it may find production distortions useful, and political realities may often dictate such an indirect policy.

3.6. *Distributional considerations*

The rules derived thus far apply to the case of identical individuals, but heterogeneity with respect to taste and ability is an important consideration. Taking account of individual differences in a population of H individuals means replacing the indirect utility function of the representative individual, $V(\boldsymbol{p}, y)$, with a social welfare function, $W(V^1(\boldsymbol{p}, y^1), \ldots, V^H(\boldsymbol{p}, y^H))$. With either fixed producer prices or constant returns to scale, there is no lump-sum income y^h, and social welfare is still simply a function of the price vector $\boldsymbol{p}$. This has the immediate implication that the production efficiency theorem just derived still holds, because there is no scope for improving social welfare once the price vector is established through the optimal tax vector $\boldsymbol{t}$. However, the shape of the social welfare function influences the choice of $\boldsymbol{t}$ itself.

The first-order conditions corresponding to maximizing this social welfare function subject to the revenue constraint (3.1) are analogous to those in Equation (3.8):

$$-\sum_h W_h \lambda^h x_i^h + \mu \left[x_i + \sum_j t_j \sum_h \frac{\mathrm{d}x_j^h}{\mathrm{d}p_i} \right] = 0, \qquad i = 1, \ldots, N, \tag{3.23}$$

where W_h is the partial derivative of W with respect to the utility of individual h, λ^h is individual h's marginal utility of income, and x_i^h is individual h's consumption of good i. Again defining $\alpha^h \equiv W_h \lambda^h + \mu \sum_j t_j \, \mathrm{d}x_j^h/\mathrm{d}y^h$ as individual h's social marginal utility of income, Equation (3.23) can be expressed in more compact form [Diamond (1975)] as

$$\sum_j t_j S_{ji} = -\frac{(\mu - \tilde{\alpha}_i)}{\mu} x_i, \qquad i = 1, \ldots, N, \tag{3.24}$$

where $S_{ji} = \sum_h S_{ji}^h$ is an aggregation of comparable terms from individual Slutsky matrices, and

$$\tilde{\alpha}_i \equiv \sum_h \left(\frac{x_i^h}{x_i} \right) \alpha^h \tag{3.25}$$

is the social marginal utility of income taken from households via a tax on good i. It is higher, the greater the share of the tax burden borne by individuals with a high social marginal utility of income, which is typically thought to be those of lower income.

Equation (3.24) is easy to understand by reference to (3.18), which still holds in this case, for $\tilde{\alpha}_i$ in place of α. Now, the marginal excess burden, rather than being equal for each source of funds, should be reduced for those commodities for which the associated loss in real income is costly ($\tilde{\alpha}_i$ is high). Because the ultimate objective is to equalize μ cross sources of revenue, those with higher distributional costs should have lower efficiency costs.

To illustrate this trade-off between equity and efficiency in the choice of tax structure, consider again the three-good case in which two consumption goods are taxed. Now, the ratio of the tax rates on the two goods should satisfy

$$\frac{\theta_1}{\theta_2} = \frac{\pi_1\varepsilon_{20} + \pi_1\varepsilon_{21} + \pi_2\varepsilon_{12}}{\pi_2\varepsilon_{10} + \pi_1\varepsilon_{21} + \pi_2\varepsilon_{12}}, \tag{3.26}$$

where $\pi_i \equiv (\mu - \tilde{\alpha}_i)/\mu$. Here, $\theta_1 > \theta_2$ if and only if $\varepsilon_{10}/\varepsilon_{20} < \pi_1/\pi_2$. If the good most complementary with leisure is also the good with the greater social valuation $\tilde{\alpha}_i$, it is not clear which good will be taxed more heavily – the answer depends in part on the strength of distributional preferences.

If preferences satisfy the restriction of homothetic separability mentioned above in Section 3.4, it will still be true that commodity taxes should be uniform (as long as preferences over consumption are the same across individuals). When preferences take this form, Engel curves (relating consumption to income) are linear and pass through the origin. Thus, there will be no variation in the relative budget shares of different goods among individuals of different abilities, and hence nothing to be gained from a distributional perspective by imposing differential taxation; this leaves the optimality of uniform taxation undisturbed.

An instance in which distributional preferences necessarily work in the opposite direction of minimizing excess burden is that in which the social welfare is the sum of individual utilities and individuals have identical Stone–Geary utility functions of the type considered in the example above, differing only with respect to ability (as measured by the wages received per unit of labor supplied). To see this, note first that the ordinary demand functions $x_i(\boldsymbol{p}, y)$ are linear in income. Thus, the change in tax revenue generated when a household changes its consumption in response to receiving a dollar of income is constant across households. This implies that differences in $\tilde{\alpha}_i$ arise only from differences in consumption patterns of households with differing social marginal utilities of income ($W_h\lambda_h = \lambda_h$). Next, note that the derivative of good-i consumption with respect to household utility is $\mathrm{d}x_i^c(\boldsymbol{p}, U)/\mathrm{d}U = (x_i - a_i)/U$, so that the elasticity of x_i with respect to U is $(x_i - a_i)/x_i$. Thus, the good with the higher elasticity of consumption with respect to utility – the good more concentrated among higher-utility individuals and hence with the lower value of $\tilde{\alpha}_i$ – is the good with the lower value of a_i relative to x_i and therefore a higher demand

cross-elasticity with respect to leisure. Thus, the good that is desirable to tax more heavily for distributional reasons is also the good that is desirable to tax less heavily for efficiency reasons.

4. Income taxation

4.1. Linear income taxation

In analyzing taxes on a representative individual, it was convenient to side-step the question of why the government might not be able to use lump-sum taxes. With population heterogeneity now an explicit aspect of the analysis, it is appropriate to revisit this question. In practice, governments include *uniform* lump-sum taxes among their tax instruments. Indeed, the use of lump-sum taxes permits the introduction of the most rudimentary of progressive income taxes, the *linear income tax*. For example, in the three-good case considered earlier, with the household's budget constraint given by Equation (3.7) and suitably modified by introducing a lump-sum tax and choosing one of the consumption goods (good 1) as the untaxed numeraire commodity, the household faces the budget constraint

$$q_1 x_1 + \frac{q_2}{1-\theta_2} x_2 = -T + \frac{w}{(1-\theta_0)} L = wL - (T + \tau wL), \tag{4.1}$$

where $\tau = -\theta_0/(1-\theta_0)$ is the household's marginal income tax rate. As Equation (4.1) shows, the government has the option of using differential commodity taxation to supplement the linear income tax schedule. This leads immediately to two questions. First, when will the government wish to use the commodity tax θ_2 or, for the case of several commodities $1, \ldots, N$, the commodity taxes $\theta_2, \ldots, \theta_N$? Second, under what conditions will the income tax be progressive, with average tax rates rising with income (e.g., with $T < 0$)?

In answer to the first question, a sufficient condition for the optimality of uniform commodity taxes or, equivalently, taxes only on labor income, is that preferences are weakly separable into goods and leisure, and that commodities have linear Engel curves with identical slopes across households [Deaton (1979)][17]. Such preferences include the case of homothetic separability, for which Engel curves pass through the origin. It is noteworthy that this condition is the same as that required for exact aggregation of consumers, and that for an aggregate measure of excess burden to be independent of the distribution of resources across consumers. Note also that a weaker condition suffices with a nonlinear income tax schedule, the design of which is discussed below. In that case, it is possible to dispense with the requirement that Engel curves be linear, since weak separability of goods and leisure suffices [Atkinson and Stiglitz (1976)].

[17] An example is the Stone–Geary utility function considered above.

If the government taxes only labor income, then Equation (3.25) implies (because purchases of labor are negative) that

$$t_0 S_{00} = -\frac{(\mu - \tilde{\alpha}_0)}{\mu}(-L), \tag{4.2}$$

where L and S_{00} are aggregate measures, with labor measured in efficiency units so that it is possible to aggregate over individuals of different abilities. The availability of lump-sum taxes adds a marginal condition that $\mu = \bar{\alpha}$, the unweighted average value of α across individuals: since the government can use positive or negative lump-sum taxes at the margin, the marginal cost of funds must equal the cost of raising funds with lump-sum taxes. Substituting this condition into Equation (4.2) and rearranging terms yields

$$\frac{(-t_0)}{p_0}\frac{p_0(-S_{00})}{L} = -\frac{(\tilde{\alpha}_0 - \bar{\alpha})}{\bar{\alpha}} \tag{4.3}$$

which, for a household labor price of $p_0 = w(1 - \tau)$ and $t_0 = -\tau w$ (recall that in this notation a positive value of t_0 raises the after-tax wage rate) may be expressed [Dixit and Sandmo (1977)] as

$$\frac{\tau}{(1-\tau)} = -\frac{(\tilde{\alpha}_0 - \bar{\alpha})}{\bar{\alpha}\bar{\varepsilon}} = -\frac{\operatorname{cov}\left(\frac{L^h}{\bar{L}}, \frac{\alpha^h}{\bar{\alpha}}\right)}{\bar{\varepsilon}}, \tag{4.4}$$

where $\bar{\varepsilon} \equiv w(1 - \tau)(-S_{00})/L$ is the aggregate compensated labor supply elasticity (which must be positive), L^h is household h's labor supply, and $\bar{L}$ is the average value of L^h across households. Since labor is expressed in efficiency units (at the common wage w), higher ability translates, for a given fraction of time worked, into higher labor supply. Expression (4.4) says that the marginal tax rate on labor income is positive if and only if the marginal social valuation of income falls as labor supply (in efficiency units) rises, a condition that is met by utilitarian social welfare functions together with labor-supply schedules that are increasing in ability.

The value of the marginal tax rate, and whether it is sufficiently high to make the linear income tax progressive ($T < 0$), depends on the weight of the social welfare function's redistributive component – how fast α^h declines as L^h rises. Properties of the marginal tax rate also depend on the amount of tax revenue required. To understand why, consider the case in which the government's revenue requirement is zero. Then it is possible to obtain a Pareto optimum by setting the marginal income tax rate, and the lump-sum tax T, to zero. Since the social marginal utility of income differs across individuals, and since there is no first-order excess burden from the introduction of a small tax, it must then be optimal to introduce some distortion (i.e., a positive marginal tax rate) to redistribute income from those with high incomes and low social marginal utility of income to those with lower incomes and higher

social marginal utility of income. Thus, the linear income tax is progressive at zero net revenue. As the government's revenue requirement rises, holding T constant, the marginal excess burden of raising revenue also rises, and so too does the cost of redistribution. As Stiglitz (1987) notes, there exists a point at which maximum revenue is collected via marginal tax rates (i.e., the marginal excess burden per dollar of revenue is infinite), at which point the government *must* rely on lump-sum taxes for additional revenue. Greater reliance on lump-sum taxes obviously reduces the progressivity of the tax schedule. Indeed, simulations confirm that the lump-sum transfer falls as revenue rises [Stern (1976)], and that it becomes negative for sufficiently high revenue requirements [Slemrod et al. (1994)].

4.2. Nonlinear income taxation: introduction

In practice, governments use income tax systems with multiple marginal tax rates. Although the linear income tax just considered can have progressive average tax burdens, its redistributive potential is limited by the fact that the average tax burden must approach the marginal tax rate asymptotically and can rise no higher. Historically, many in government have felt that only a schedule of rising marginal tax rates could deliver the appropriate degree of progressivity toward the top of the income distribution, and have implemented income tax systems with top marginal tax rates in some instances exceeding 90 percent[18].

Governments certainly can impose income tax systems more complicated than the linear income tax, but what should these systems look like? As in the case of the linear income tax, the issue involves balancing efficiency and equity, with the surprising conclusion that high and rising marginal tax rates may well not be appropriate even when the government has a strong redistributive motive.

At first, it might seem that the ability to choose an arbitrary income tax function $T(\cdot)$ offers the government the opportunity to impose individual-specific lump-sum taxes, for the function could be chosen to pass through values of tax burdens appropriate to individuals at each level of income. However, as is rapidly apparent, the endogeneity of income strongly limits the government's ability to impose differential lump-sum taxation.

To begin, suppose that there is a single consumption good, that labor supply is the only source of income, and that individuals have common preferences $U(c, l)$ over consumption and leisure, differing only in their abilities, as measured by wage rates w. Imagine that the government needs to raise a certain amount of revenue, R, using an income tax, and that it is desirable to assign a lump-sum income tax burden T_i to

[18] For example, just prior to the Kennedy–Johnson tax cut of 1964, the top marginal federal income tax rate in the United States was 91 percent.

individual i. With the consumption good as numeraire, the problem may be expressed as

$$\max_{T} W(V^1(w^1, -T^1), V^2(w^2, -T^2), \ldots, V^H(w^H, -T^H)) \text{ subject to } \sum_h T^h \geqslant R. \tag{4.5}$$

If μ is the Lagrange multiplier associated with the revenue constraint, then the H first-order conditions are simply that $W_h \lambda^h = \mu$ – that the marginal social utility of income is the same across all individuals.

What does this condition imply for tax burdens? For the utilitarian social welfare function $W(U^1, \ldots, U^H) = \sum_h U^h$, it implies that the marginal utility of income λ^h is constant across individuals, which (from the first-order conditions for utility maximization) implies that the marginal utility of consumption is constant across households, but that the marginal utility of leisure is proportional to w^h. Equating the marginal social cost of *income* across individuals, the government in effect forces high-wage individuals to work until they reach the point that leisure is very valuable to them. In the process, this tax system makes high-wage individuals *worse off* than low-wage individuals, a paradoxical outcome that is guaranteed if leisure is a normal good.

For example, suppose the common utility function takes the quasi-linear form $U(c, l) = c - v(1 - l)$, with $v' > 0$ and $v'' > 0$. Then, with optimal household-specific taxation, all households have the same level of consumption, and leisure declines monotonically with the wage rate. The lowest-wage household obtains the highest level of utility, which illustrates quite clearly the problem to be faced in attempting to implement such a tax system. Aside from the political implausibility of the outcome, this scheme could be implemented only if government knew each household's ability level and assigned taxes accordingly. Otherwise, all other households would have incentives simply to masquerade as the household with the lowest ability by supplying the amount of labor necessary to produce that household's income level, thereby leaving themselves better off than the lowest-ability household (because they forgo less leisure to reach this level of income), rather than worse off. But this, in turn, leaves the government with a uniform lump-sum tax and too little revenue. While the government could respond by increasing the lump-sum tax, it is clear from the previous discussion of the linear income tax that this policy alone is not likely to be optimal. Rather, the government seeks to impose a tax system more progressive than the lump-sum tax, while still accounting for the absence of information about individual types and the endogeneity of household income. A linear income tax is but one such tax system.

4.3. Nonlinear income taxation: graphical exposition

Much of the intuition behind the design of the optimal nonlinear income tax emerges from consideration of an income tax imposed on an economy composed of two

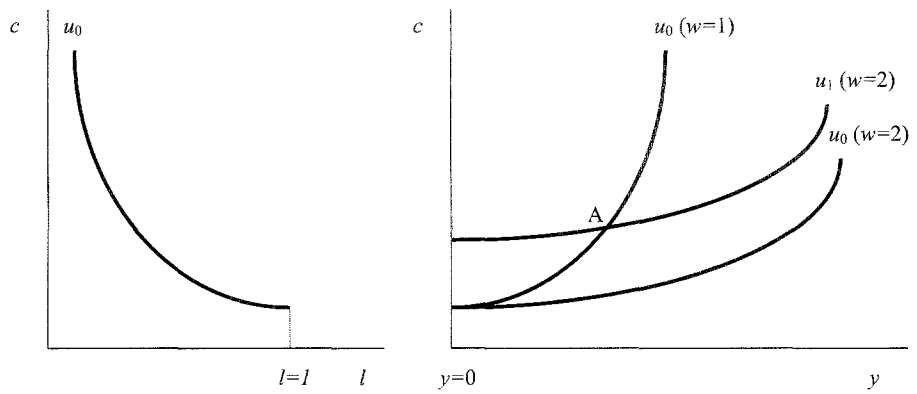


Fig. 7. Indifference curves over consumption and income.

individuals, one (H) of high ability and one (L) of low ability[19]. Because the government observes only income, $Y = w(1 - l)$, rather than labor supply and wage rates separately, it is useful to express each individual's preferences over consumption and leisure (or labor) in terms of preferences over consumption and income, as depicted in Figure 7. On the left-hand side of the figure is an indifference curve over consumption and leisure, based on the utility function $U(c, l)$. On the right are two corresponding indifference curves for the same level of utility but different wage rates, based on the same utility function, $U(c, 1 - y/w)$. The curve corresponding to the higher wage rate is flatter because a given change in labor translates into a greater change in income. This suggests that when indifference curves of two individuals do cross, as at point A, the indifference curve of the higher-ability individual is flatter.

Figure 8 illustrates the outcome of attempting to impose the previously discussed lump-sum tax "solution", with consumption equal to c_0 for both high- and low-ability individuals and the higher-ability type on a lower indifference curve, as indicated by the relative consumption at zero income (at which ability differences are irrelevant). Rather than accept the bundle (c_0, y^H), the high-ability household would prefer to earn income y^L and receive the same level of consumption. The problem with this plan is that it violates the *self-selection constraint* that each household prefer its government-designated bundle among the available options. In this instance, the high-ability household prefers the bundle designated for the low-ability household. It is typically the self-selection constraint of the high-ability person with which the government must be concerned.

As Figure 9 illustrates, the self-selection constraints limit the scope for redistribution through differential lump-sum taxation. For the sake of exposition, assume that the

[19] We follow the mnemonic notation in the literature in denoting the two ability classes as H and L for the following graphical exposition, but remind the reader that the variable L represents labor supply in all other parts of the chapter.

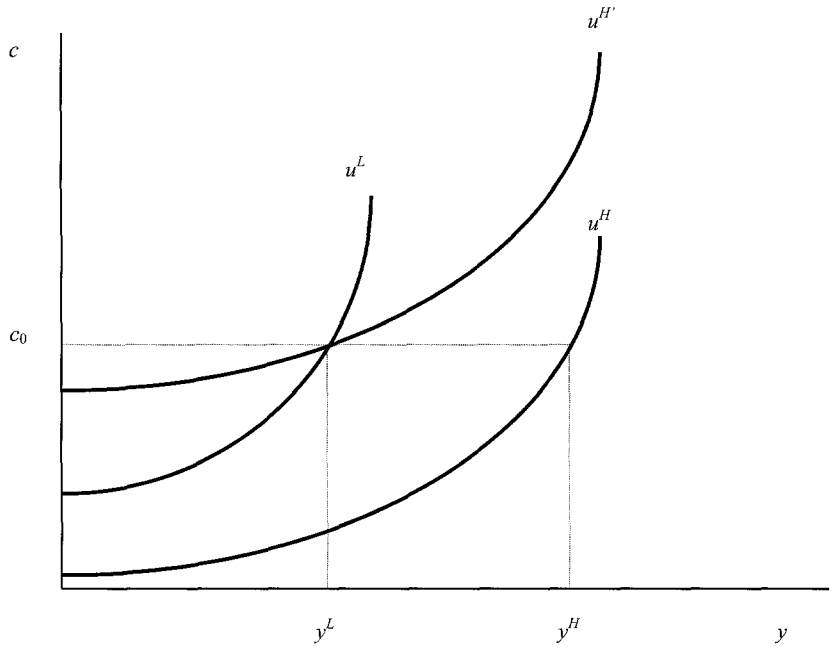


Fig. 8. Violation of the self-selection constraint.

required level of revenue, R, equals zero. With no redistribution, each household's budget constraint has unit slope (since a dollar of income produces a dollar of consumption) and passes through the origin. The high-ability and low-ability households choose points H and L, respectively. Each household strictly prefers its own bundle, so neither self-selection constraint is binding. As a result, it is possible to impose a lump-sum tax on H and provide an equal lump-sum transfer to L until reaching the point that H's self-selection constraint binds, which occurs at points H′ and L′. The government cannot do more with lump-sum taxation *without* violating H's self-selection constraint, but it *can* do more.

Slopes of the indifference curves of individuals H and L differ at point L′. Because this point is an optimum for L (since L's indifference curve is tangent to the budget line) but not for H, a slight movement in any direction along the budget line has no first-order effect on the utility of L, but does have a first-order effect on the utility of H. Moving toward the origin along the budget line makes H worse off, because H is already working inefficiently "too little" at point L – H's indifference curve is flatter than the budget line. This suggests a way to relax H's self-selection constraint and achieve more redistribution, as illustrated in Figure 10. By shifting individual L from point L′ to point L″, the government imposes on L only a "second-order" excess burden (since L is initially at an undistorted point) but raises "first-order" tax revenue

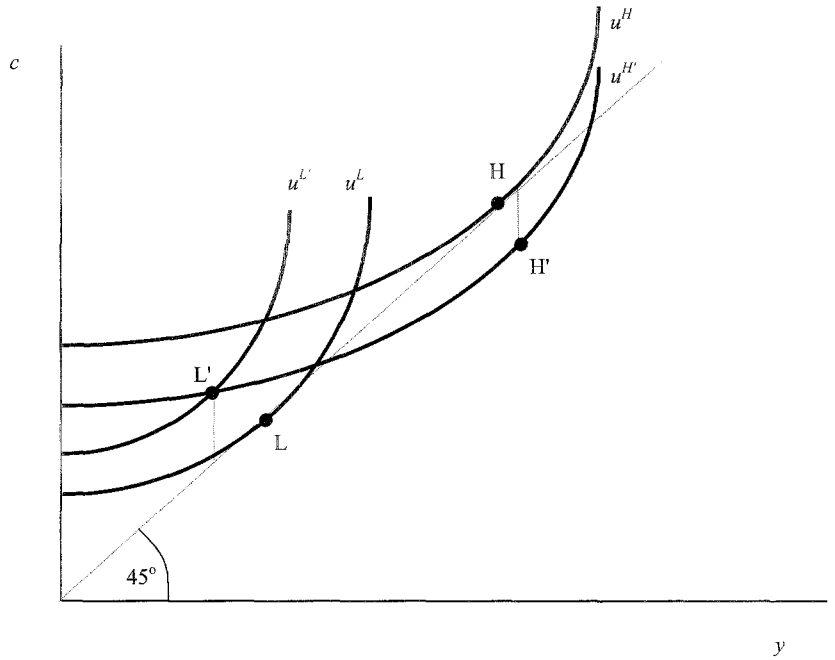


Fig. 9. The scope for lump-sum taxation.

by being able to shift individual H down to point H''. This tax revenue equals the distance CD in Figure 10. The extra revenue extracted from H (net of the amount – distance AB in Figure 10 – needed to compensate for the small distortion to L's choice) can then be allocated between L and H, with H receiving just enough to keep the self-selection constraint satisfied. The final result is that L is better off than at L' and H is worse off than at H'.

The limits that govern this redistribution are the government's success in carrying it out (which reduces disparities in the social valuation of marginal incomes received by different households) and by marginal excess burdens that rise as one moves further away from the initial point L'. L's bundle can be thought of as being implemented via a marginal tax rate on L's income that produces a budget line with slope less than one. This offers the insight that it is optimal to impose a positive marginal tax rate on individual L not to raise revenue from L, but to raise revenue from those with incomes higher than L's – in this case, individual H. A corollary is that, as there is no one of higher ability than H in this example, it is not optimal to impose a marginal tax rate on H's income. Doing so would distort H's behavior and reduce the revenue the government could extract from H without violating H's self-selection constraint. These lessons are useful in considering the case in which there is a continuum of agents.

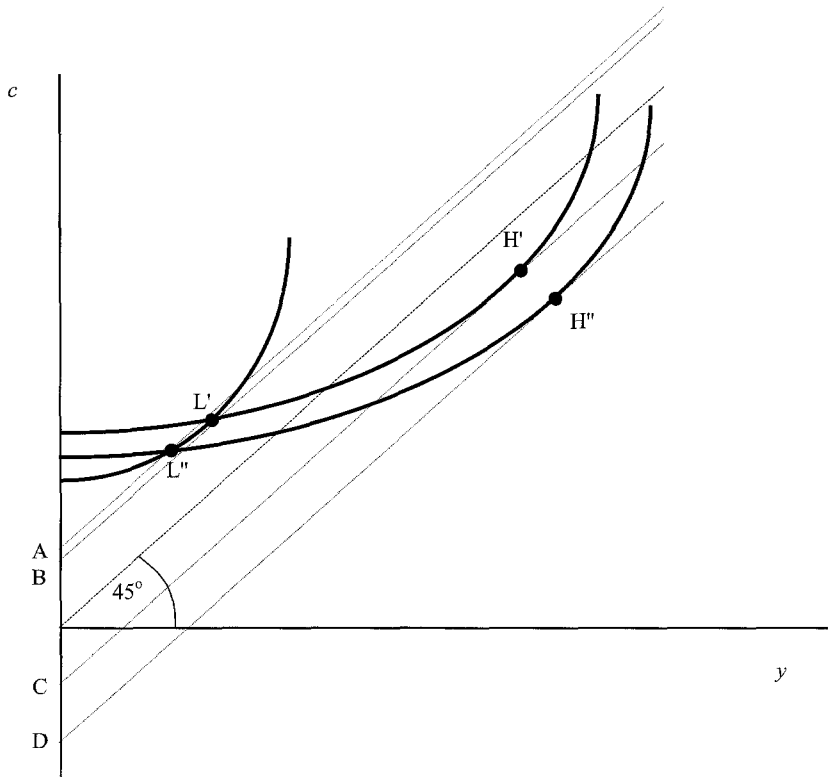


Fig. 10. Using distortionary income taxation.

4.4. Nonlinear income taxation: mathematical derivation

The mathematics of optimal income taxation with a continuum of agents is not straightforward, because it is not possible to rule out such phenomena as nondifferentiability of the tax function $T(\cdot)$. These phenomena are not simply "anomalies". As discussed in Stiglitz (1987), nondifferentiability arises in cases in which it is optimal to pool individuals with different skill levels at a single point in (c,y) space. To understand why, consider the case in which there are many individuals of type H (as considered above) and an equal number of individuals of type L. The optimal tax policy is obviously identical to that with one individual of each type. Then introduce an additional individual at some intermediate wage rate between L and H. If this individual, say M, is offered an allocation that H prefers to L's bundle, then H's self-selection constraint is violated. It is possible to maintain individuals of type H at their initial allocations only by reducing the attractiveness of M's bundle. This, in itself, distorts the choice of M's bundle, but if there are many more individuals of types H and L than of type M, society gains from doing so until M's bundle approaches that of L.

In spite of the importance of this complication, it is useful for intuition to derive results for cases in which such problems do not arise. Our approach closely follows that in Atkinson and Stiglitz (1980). For further discussion of the more general mathematical issues, see Mirrlees (1976, 1986).

Continuing to assume, for simplicity, that overall revenue $R=0$, the government seeks to maximize some general social welfare function of individual utilities, subject to the constraint that total consumption equal total before-tax income. Letting $f(w)$ be the fraction of the population endowed with wage rate/skill level w, the government's objective is

$$\max \int_w G(U(w))f(w)\,\mathrm{d}w \quad \text{subject to} \quad \int_w (c(w)-y(w))f(w)\,\mathrm{d}w \leqslant 0, \tag{4.6}$$

where $c(w)$ and $y(w)$ are the levels of consumption and income chosen by each individual at wage rate w, and $U(w)$ is the utility of that individual based on these values, $U(c(w), 1-y(w)/w)$.

The optimization problem is further constrained by the requirement that wage-w individuals voluntarily choose the bundle $(c(w), y(w))$ – the self-selection constraint discussed above. The requirement that the bundle $(c(w), y(w))$ is individually rational for people of wage w means that utility $U(c(w'), 1-y(w')/w)$ achieves a maximum at $w'=w$. This may be expressed in terms of the first-order condition

$$\frac{\partial U}{\partial c}\frac{\mathrm{d}c}{\mathrm{d}w'}+\frac{\partial U}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}w'}=0 \tag{4.7}$$

that indicates that the individual cannot increase utility through a local change in labor supply. This then implies, for common preferences, that the change in utility as the wage rate rises is simply the derivative of the utility function with respect to w, holding c and y fixed:

$$\frac{\mathrm{d}U}{\mathrm{d}w}=\frac{\partial U}{\partial w}=U_2\frac{y}{w^2}=U_2\frac{L}{w}. \tag{4.8}$$

Thus, the optimal tax problem is that expressed in Equation (4.6), subject to the additional constraint given in Equation (4.8). While it is expressed as one of choosing the bundle (c,y), it can equally well be viewed as a choice of the utility level u and the level of labor supply L, as $u=U(c,1-L)$ and $y=wL$. To solve the problem expressed this way, it is helpful to form the Hamiltonian

$$H=[G(u)-\mu(c(L,u)-y(L,u))]f(w)-\eta U_2(L,u)\frac{L}{w} \tag{4.9}$$

with control variable L, state variable u, Lagrange multiplier μ and costate variable η. The first-order conditions are

$$\text{(a)}\ \frac{\partial H}{\partial L}=0, \qquad \text{(b)}\ \frac{\partial H}{\partial u}=-\frac{\mathrm{d}\eta}{\mathrm{d}w}. \tag{4.10}$$

Condition (4.10a), as applied to Equation (4.9), implies that

$$-\mu\left[\left.\frac{\partial c}{\partial L}\right|_u - \left.\frac{\partial y}{\partial L}\right|_u\right] f(w) - \eta\left[\left.\frac{\partial U_2}{\partial L}\right|_u \cdot \frac{L}{w} + \frac{U_2}{w}\right] = 0. \tag{4.11}$$

Note that

$$y = wL \Rightarrow \left.\frac{\partial y}{\partial L}\right|_u = w \quad \text{and} \quad \left.\frac{\mathrm{d}u}{\mathrm{d}L}\right|_u = 0 = U_1 \left.\frac{\partial c}{\partial L}\right|_u - U_2 \Rightarrow \left.\frac{\partial c}{\partial L}\right|_u = \frac{U_2}{U_1}.$$

Further, individual utility maximization ensures that $U_2/U_1 = w(1 - T')$. Thus, Equation (4.11) can be rewritten as

$$\frac{T'}{1 - T'} = \left(\frac{U_1 \eta}{\mu}\right) \frac{\psi}{wf(w)}, \tag{4.12}$$

where $\psi \equiv \left.\frac{\partial U_2}{\partial L}\right|_u \cdot \frac{L}{U_2} + 1$. This expression says that the optimal marginal tax rate is increasing in $(U_1\eta/\mu)$ and ψ and decreasing in $wf(w)$. The last of these effects is straightforward: the more effective labor supply that is subject to the marginal tax rate at w, the greater is the excess burden associated with that tax rate.

To interpret the other two terms in Equation (4.12) and their effects, consider the special case of quasilinear preferences, $U(c, l) = c - v(1 - l) = c - v(L)$, where $v(\cdot)$ is convex. For this case, it may be shown that $\psi = 1 + 1/\varepsilon$, where ε is the compensated labor-supply elasticity at w. Thus, a higher labor-supply elasticity leads to a lower value of ψ, which by Equation (4.12) leads to a lower marginal tax rate. This is sensible, as a higher labor-supply elasticity is also associated with greater excess burden per dollar of revenue raised. A similar effect appears in Equation (4.4) for the case of the linear income tax, but here it is the labor-supply elasticity at the particular wage rate w, rather than the aggregate labor-supply elasticity, that is important because the government is free to choose different marginal tax rates for different levels of income.

Finally, consider the remaining term in Equation (4.12), $(U_1\eta/\mu)$. From the first-order condition (4.10b),

$$-\frac{\mathrm{d}\eta}{\mathrm{d}w} = \frac{\partial H}{\partial u} = \left[G' - \mu\left(\left.\frac{\partial c}{\partial u}\right|_L - \left.\frac{\partial y}{\partial u}\right|_L\right)\right] f(w) - \eta U_{21} \frac{L}{w} \left.\frac{\partial c}{\partial u}\right|_L. \tag{4.13}$$

As $\partial y/\partial u|_L = 0$ and because $\mathrm{d}u/\mathrm{d}u|_L = 1 \Rightarrow \mathrm{d}c/\mathrm{d}u|_L = 1/U_1$, Equation (4.13) can be rewritten as

$$\frac{U_1}{\mu}\frac{\mathrm{d}\eta}{\mathrm{d}w} = -\left[\frac{G'U_1}{\mu} - 1\right] f(w) + \frac{\eta U_{21} L}{w\mu}. \tag{4.14}$$

To interpret this further, it is again helpful to impose the simplifying assumption of quasilinear preferences, thereby implying that U_1 is constant (here normalized

to 1) and $U_{21}=0$. Then, integrating both sides of Equation (4.14) and imposing the transversality condition ($\eta \to 0$ as $w \to \infty$) yields

$$\frac{U_1 \eta}{\mu} = \int_w^\infty \left(1 - \frac{G'(\tilde{w})\,U_1}{\mu}\right) f(\tilde{w})\,\mathrm{d}\tilde{w} = [1 - F(w)] - \int_w^\infty \frac{G'(\tilde{w})\,U_1}{\mu} f(\tilde{w})\,\mathrm{d}\tilde{w}, \tag{4.15}$$

where $F(\cdot)$ is the cumulative density function based on $f(\cdot)$[20].

This expression equals the social value, scaled by the marginal cost of funds μ, of raising a dollar through marginal taxation at wage level w. This value has two components. The first term is the amount of revenue raised, equal to the taxes collected from all those who pay the extra tax – those with wage rates at least as high as w. The second term is the value, again in revenue units, of the social welfare lost by these individuals in paying the extra tax. Each of these terms declines with w, because we collect less revenue and impose less burden by raising taxes on fewer people, but it is the difference between the terms that matters. What pattern does this difference follow? The difference must be positive if marginal tax rates are positive, and the difference converges to zero as $w \to \infty$. If G' declines with w, then the second term in Equation (4.15) – the social cost of an increase in the marginal tax rate at w – converges to zero more rapidly than does the first term. Hence, there may be a range of w over which the difference between the two terms increases. The intuition is that high marginal tax rates at high levels of income are very inefficient because they produce so little revenue, while high marginal tax rates at low levels of income are inequitable because they impose burdens on those with very high social marginal utilities of income G'. The best compromise may be to raise marginal tax rates at middle income levels, where tax obligations are not imposed on those for whom the burden of higher taxes is most socially costly but where higher tax rates still raise considerable revenue.

As should be clear from this discussion, the exact pattern that the term in Equation (4.15) follows as w rises depends on the social welfare function and the shape of the wage distribution. Even if this term does indicate higher marginal tax rates somewhere in the middle of the wage distribution, this is precisely where one of the other terms in Equation (4.12), $wf(w)$, is also likely to be greatest, which has the effect of reducing T'.

Thus, it is possible to say very little about the general shape of the optimal marginal tax rate schedule, although since the seminal work of Mirrlees (1971) there has been a general tendency to find that optimal marginal tax rates should either fall throughout most of the income distribution or else have an "inverted-u" shape, reflecting the effect of the term in Equation (4.15) [see, e.g., Kanbur and Tuomala (1994)]. This conclusion

[20] In recent work, Saez (2000a) derives an analytical expression extending Equation (4.15) to the case of more general preferences. While he offers an intuitive discussion of this expression, it is necessarily more complicated than the basic intuition presented here.

is in some sense predetermined by findings that, under certain circumstances, the optimal marginal tax rate equals zero at both the top and bottom of the income distribution.

The rationale for a zero top marginal tax rate appeared already, in the graphical presentation of the two-person case. For the general case with a bounded distribution of wage rates, the result [see Phelps (1973), Sadka (1976) and Seade (1977)] follows directly from the fact that the term in Equation (4.15) approaches zero as the wage w approaches its upper support, $\bar{w}$. As to why the marginal rate might be zero at the bottom of the wage distribution [see Seade (1977)], consider the value of expression (4.15) at the lower support of the wage distribution, say $\underline{w}$. As $F(\underline{w}) = 0$, the expression indicates that $T'/(1-T') \propto 1 - \bar{\alpha}/\mu$, where $\bar{\alpha}$ is the average social marginal utility of income over the entire distribution[21]. But, as discussed in the case of the linear income tax, $\bar{\alpha} = \mu$ when there is a uniform lump-sum tax available, so T' must equal zero. The intuition for this result follows the algebra. At the very bottom of the income distribution, an increase in the marginal tax rate has the same revenue and distributional effects as a uniform lump-sum tax – it raises revenue from the entire population. But it also distorts the behavior of the lowest-income individuals, which a lump-sum tax does not. Thus, a lump-sum tax dominates any positive marginal tax on lowest-wage individuals.

However, neither of these results is robust to reasonable changes in assumptions. As its derivation suggests, the result regarding the marginal tax rate at the bottom requires that the entire population works. Otherwise, the marginal tax rate applied to the lowest-wage *worker* does not collect tax revenue from all individuals, and the logic just given breaks down[22]. At the top of the wage distribution, optimal marginal tax rates need not approach zero, even in the limit, if the wage distribution is unbounded, nor is the "inverted-u" shape of the marginal tax rate distribution robust, as demonstrated by Diamond (1998) for the case of a Pareto distribution of wages and quasilinear preferences[23].

Even for bounded wage distributions where optimal marginal tax rates must eventually decline, marginal tax rates may rise over most of the income distribution,

[21] As there are no income effects on labor supply for the quasilinear utility function, it is possible to ignore the indirect effect of income on revenue.

[22] A different departure from this logic occurs if individuals at the bottom end of the income distribution make discrete choices of whether or not to work, as analyzed by Saez (2000b). In this case, the optimal marginal tax rate on the lowest income is negative, since the tax system thereby induces greater labor-force participation and higher incomes.

[23] Diamond finds the optimal marginal tax rate schedule to be u-shaped in the example he analyzes. As clarified by Dahan and Strawczynski (2000), though, Diamond's result of a rising marginal tax rate at the top depends on the joint assumptions of an unbounded ability distribution *and* quasilinear preferences. The result need not hold, even for the Pareto distribution of abilities, if one adopts a more general utility function. For another variation in assumptions, Stiglitz (1982) notes that if the effort of high-skilled workers is an imperfect substitute for that of low-skilled workers, it may be optimal to *subsidize* income at the top of the wage distribution to increase skilled labor effort and thereby raise the wages of the less skilled.

although numerical simulations of the more restricted optimal two-bracket linear tax system [Slemrod et al. (1994)] find that the second/top marginal rate is lower than the first. This has quite interesting implications for the recent debate about the equity effects of the *flat tax* [Hall and Rabushka (1995)], a close relative of the linear income tax under which tax liabilities are constrained to be nonnegative. Although some [e.g., Bradford (1986)] have suggested modifying the flat tax to permit additional, higher marginal tax rate brackets on higher-wage individuals, these simulation results suggest that adding an additional bracket should occasion lower, not higher marginal tax rates at higher wage levels.

5. Externalities, public goods, and the marginal cost of funds

The analysis to this point ignores the use to which public funds may be put, other than redistribution to other taxpayers. In reality, of course, a major reason for raising revenue is to finance public expenditures, and it is important to consider how this affects the conclusions. In turn, it is interesting to ask how the use of distortionary taxation influences the optimality conditions of Samuelson (1954) regarding the provision of public goods. At the same time, it is convenient to consider how the distortionary nature of taxation alters the prescriptions concerning the use of Pigouvian taxation to correct externalities.

Basic results relating the provision of public goods and the correction of externalities to the use of distortionary taxes may be found, respectively, in Atkinson and Stern (1974) and Sandmo (1975). Auerbach (1985) presents and interprets these results in some detail, so we will offer only a brief derivation here. Both models assume that the government is limited to the use of indirect proportional taxes, and avoid any discussion of distribution by assuming that individuals are identical, i.e., that the population consists of H copies of the representative individual. In this context, it is natural to assume that the government seeks to maximize the utility of each representative individual or, equivalently, the sum of individual utilities.

5.1. *The provision of public goods and the marginal cost of public funds*

Consider first the case in which the government wishes to provide a public good, G, using all its tax revenue. Individuals choose consumption $\boldsymbol{x}$ treating G as given, so their utility function may be written in semi-indirect form as $V(\boldsymbol{p}, y; G)$, with $\partial V/\partial G = \partial U/\partial G|_{x(\boldsymbol{p}, y; G)}$. For simplicity, the economy's production function $f(\boldsymbol{X}, G)$ (where $\boldsymbol{X} = H\boldsymbol{x}$) is taken to obey constant returns, so that there are no pure profits and $y = 0$. This set-up gives rise to the Lagrangean (compare to 3.12):

$$HV(\boldsymbol{p}; G) - \mu f(\boldsymbol{X}, G), \tag{5.1}$$

with first-order conditions with respect to each price and the level of public goods, G. The first-order conditions with respect to price are identical to those derived in Section 3 for the case of $y = 0$, in Equation (3.15), repeated here for convenience:

$$-\lambda X_i + \mu \left[X_i + \sum_j t_j \frac{\mathrm{d}X_j}{\mathrm{d}p_i} \right] = 0, \qquad i = 1, \ldots, N, \tag{5.2}$$

except that X_i is now the sum of individual purchases of good i, equivalently the product of H and the purchase of the representative consumer. The first-order condition with respect to the public good is

$$H \frac{\partial V}{\partial G} - \mu \left[f_G + \sum_i f_i \frac{\partial X_i}{\partial G} \right] = 0. \tag{5.3}$$

The utility function implies that $\partial V/\partial G = U_G^h$, in which U_i^h is individual h's marginal utility of good i. The economy's production constraint and private production efficiency impose the condition that $f_i \propto q_i$, while the consumer's budget constraint implies that $\boldsymbol{p}' \partial \boldsymbol{X}/\partial G = 0$. Taking good 0 to be the untaxed numeraire commodity, and λ to be the marginal utility of income, it follows that $U_0^h = \lambda p_0 = \lambda f_0$, and Equation (5.3) implies

$$\sum_h \frac{U_G^h}{U_0^h} = \left(\frac{\mu}{\lambda} \right) \left(\frac{f_G}{f_0} - \frac{\mathrm{d}R}{\mathrm{d}G} \right), \tag{5.4}$$

where R is tax revenue, $\boldsymbol{t}'\boldsymbol{X}$, and the variable μ is the shadow cost of the government's revenue constraint (measured in units of utility). The ratio μ/λ, which measures the shadow price of revenue units of the numeraire, is often referred to as the *marginal cost of public funds* (MCPF), because it measures the cost of each unit of public funds, taking account of the deadweight loss from the additional taxes associated with those funds.

Expression (5.4) deviates in two respects from the Samuleson rule of equating the marginal rate of transformation, f_G/f_0, and the sum of the marginal rates of substitution, $\sum_h U_G^h / U_0^h$. First, it indicates that the implicit cost of public goods is reduced to the extent that public spending increases spending on taxed commodities, i.e., $\mathrm{d}R/\mathrm{d}G > 0$ – a point noted by Diamond and Mirrlees (1971a,b). Second, it requires that one adjust the relative price of public goods, f_G/f_0, for the MCPF, consistent with intuition provided by Pigou (1947). However, as noted by Atkinson and Stern, the MCPF as defined need not exceed 1. Recall from Section 3 that optimal taxes ensure that $\mu > \alpha$, where $\alpha = \lambda + \mu \mathrm{d}R/\mathrm{d}y$ is the "social" marginal utility of income – the value to society of giving an individual an extra unit of income, taking account of the revenue provided by induced spending on taxed goods. However, if $\mathrm{d}R/\mathrm{d}y$ is negative, then it is possible that the MCPF is equal to or even less than 1.

A simple example illustrating this possibility is provided by Ballard and Fullerton (1992). Consider the case in which the utility function is weakly separable into private and public goods, so that $dR/dG = 0$. Suppose that there are just two private goods, leisure and consumption, so that there is just one independent tax instrument, and normalize this tax instrument so that only the tax on labor income is positive. The first-order condition with respect to the price of labor – the wage rate w – is, from Equation (5.2),

$$-\lambda L + \mu\left(L - t\frac{dL}{dw}\right) = 0, \tag{5.5}$$

where L is the aggregate supply of labor and t is the tax per unit of labor supplied[24]. Defining η_{Lw} as the uncompensated labor supply elasticity and θ as the tax rate t/w, Equation (5.5) may be rewritten:

$$\frac{\mu}{\lambda} = \frac{1}{1 - \theta\eta_{Lw}} \tag{5.6}$$

from which it is obvious that the MCPF exceeds 1 if and only if the uncompensated labor supply elasticity is positive. For the "benchmark" case of Cobb–Douglas preferences, the uncompensated labor supply elasticity is zero, and the MCPF = 1[25]. Given that a zero uncompensated labor-supply elasticity lies within the range of existing estimates, this result is not simply a theoretical curiosity, and suggests that we may well err in automatically assuming that the existence of distortionary taxation raises the MCPF significantly[26].

The reason that this assumption has the potential to go wrong is that the deadweight loss of a tax system and the MCPF are two entirely separate concepts. Deadweight loss is a measure of the potential gain from replacing distortionary taxes with an efficient lump-sum alternative, and marginal deadweight loss is simply the change in this magnitude as tax revenue changes. By contrast, the MCPF reflects the welfare cost, in units of a numeraire commodity, of raising tax revenue for exhaustive government expenditure.

While this result seems simple and straightforward, much has been written on the topic of how the MCPF should be defined. Without reviewing this extensive literature

[24] The term $t\,dL/dw$ enters in expression (5.5) with a minus sign because the tax is subtracted from the wage.

[25] Ballard and Fullerton argue based on an informal survey that this outcome was generally a surprise to a group of public-finance economists.

[26] More generally, if the utility function is not separable, one may show that the Samuelson rule holds whenever the supply of labor is unaffected by the increase in spending on the public good – whenever the combined impact on L of the increase in G and the decrease in w equals zero. In this case, the marginal cost of funds as defined in Equation (5.3) is not equal to 1, but its deviation from 1 is offset by the dR/dG term.

[see, for example, the survey by Håkonsen (1998)], we note that the disagreements relate largely to terminology and questions of normalization. As an illustration [see Schöb (1997)], consider the same example (one public good, labor, and one other private good), but normalize the proportional taxes so that the tax on labor is zero. The first-order condition with respect to the price, p, of the taxed commodity, instead of Equation (5.5), would be

$$\lambda X + \mu\left(X + t\frac{\mathrm{d}X}{\mathrm{d}p}\right) = 0, \tag{5.7}$$

where X is the aggregate purchase of the commodity and t is the tax per unit of that commodity. Defining η_{Xp} as the uncompensated own-price demand elasticity and θ as t/p, Equation (5.7) can be rewritten as

$$\frac{\mu}{\lambda} = \frac{1}{1 + \theta\eta_{Xp}} \tag{5.8}$$

which says that the MCPF should exceed 1 if and only if $\eta_{Xp} < 0$ – i.e., X is not a Giffen good. Since this is a much weaker condition than that $\eta_{Lw} > 0$, it is easy to see how one might become confused, given that these conditions supposedly reflect the same underlying experiment. Indeed, when $\eta_{Lw} = 0$, $\eta_{Xp} = -1$, so $\eta/\lambda = 1/(1 - \theta)$. This apparent paradox is resolved by noting that the normalization does not affect the underlying outcome, but does change the units of (μ/λ). In the first instance, the MCPF is defined in units of the commodity; in the second, it is measured in terms of units of labor.

The impact of this difference may be understood using the standard approach of cost–benefit analysis [e.g., Harberger (1972)], that weights the costs of funds according to sources. When the labor-supply elasticity is zero, an increase in the tax on labor has no impact on the amount of labor supplied. Thus, the extra taxes that finance additional spending on the public good are absorbed fully through reduced consumption. Hence, the marginal cost of funds equals the marginal value of a unit of the commodity. Therefore, if the commodity is chosen as the numeraire, the marginal cost of funds equals 1. If labor is chosen as the numeraire, the marginal cost of funds still equals 1 unit of the commodity, but this equals $1/(1 - \theta)$ units of labor, due to the tax wedge between labor and private consumption. The equilibrium is the same regardless of normalization, but the MCPF is different. This discussion also highlights that the MCPF reflects only the presence of a distortion on one particular margin – between the public good and the numeraire. This distortion can be positive, negative or zero, independent of the presence of deadweight loss due to taxation.

5.2. Externalities and the "double-dividend" hypothesis

A similar logic applies to the analysis of externalities, as in Sandmo (1975). Suppose that, rather than there being a public good, there is an externality, E, that enters into

each person's utility function and which cannot be avoided, so that the representative individual's indirect utility function may be written $V(\boldsymbol{p}; E)$. Suppose also, for simplicity, that the externality is the product of aggregate consumption of a single good, say the good with the highest index, N. Then, the Lagrangean,

$$HV(\boldsymbol{p}; X_N) - \mu f(\boldsymbol{X}), \tag{5.9}$$

implies the following N first-order conditions with respect to the prices of goods $1, \ldots, N$ (compare 3.8):

$$-\lambda x_i + \mu \left[x_i + \sum_j t_j^* \frac{\mathrm{d}x_j}{\mathrm{d}p_i} \right] = 0, \qquad i = 1, \ldots, N, \tag{5.10}$$

where

$$t_j^* = t_j, \quad j \neq N,$$
$$t_N^* = t_N + \frac{HV_E}{\mu} = t_N + \frac{HV_E/\lambda}{\mu/\lambda}.$$

Expression (5.10) is the standard optimal tax solution, except that it calls for the tax on the externality-producing good, t_N, to equal the sum of the "optimal" tax that ignores the externality, t_N^*, plus a term that reflects the cost of the externality. This second term equals the corrective Pigouvian tax – the social cost per unit of consumption of the good, measured in terms of the numeraire commodity – divided by the MCPF, μ/λ.

Thus, in a result analogous to that just presented for the provision of public goods, the presence of distortionary taxation leads to "undercorrection" of the externality if and only if the MCPF exceeds 1. As before, though, one must exercise care in interpreting this result. Suppose, following the previous example, that the externality enters the utility function in a separable manner, and that preferences over direct consumption of goods and leisure are Cobb–Douglas. Also assume that there are just two consumption goods, a "clean" good and a "dirty" good that causes the externality. Absent the externality (and if various regularity conditions are satisfied), the optimal tax structure calls for equal taxes on the two consumption goods, i.e., $t_1^* = t_2^*$. This can be achieved either through a tax on wages alone or through uniform taxes on the two consumption goods. In the first case, letting the clean good be numeraire, it is clear that $\mu/\lambda = 1$, so the Pigouvian tax should be implemented without adjustment. In the second case, letting labor be numeraire, $\mu/\lambda = 1/(1-\theta) > 1$, so it is necessary to "undercorrect" for the externality.

It is tempting to conclude in the latter case that one "undercorrects" because the corrective tax is piled on top of the pre-existing consumption tax, while in the former case no initial pre-existing consumption tax exists. However, the two equilibria are

identical, with the same distortions present on all margins[27]. Thus, the intuition is misleading. While there is no initial consumption tax when only labor is taxed, there is still a distortion of the labor–leisure choice. Taxing the dirty consumption good exacerbates the distortion between that good and labor, just as if the initial tax were on the two consumption goods instead. The fact that it is overall distortions that matter, and not the levels of individual taxes, also exposes a serious interpretive difficulty in what is known as the "double-dividend" hypothesis. This hypothesis, as discussed in much more detail in chapter 23 of this Handbook by Bovenberg and Goulder, states that corrective taxes have an added benefit in the presence of other distortionary taxes – the revenue that allows a reduction in the other tax rates and their associated deadweight loss. Corrective taxes do not merely raise revenue and correct externalities, but also exacerbate existing distortions. Taxing consumption and using the proceeds to reduce taxes on labor has no net impact on the consumption–leisure choice in this instance.

5.3. Distributional considerations and the MCPF

With a heterogeneous population, the provision of public goods and the correction of externalities take on added complications. Even in the absence of distortionary taxation, the optimal rules then reflect the social valuations of utilities of different individuals. In addition, the costs and benefits of public goods, externalities, and the taxes that address them all have distributional consequences. For example, the government might wish to expand provision of public goods that have favorable distributional consequences; Sandmo (1998) offers a detailed analysis of the general problem. Also see Slemrod and Yitzhaki (2001), who illustrate how one can decompose both the costs and benefits of public expenditure projects in terms of efficiency and distributional consequences. However, it is also useful to consider circumstances in which the problem becomes much simpler, which is the case when the government has sufficient flexibility in its choice of tax instruments.

There is a close analogy here to the standard optimal income tax problem, under which it may not be necessary to tax luxury goods more heavily for purposes of distribution if the government can use a nonlinear income tax [as in Atkinson and Stiglitz (1976)]. Indeed, the analysis yields a parallel result, namely that distributional considerations should not enter into the provision of public goods or the correction of externalities when there is a nonlinear income tax and preferences are weakly separable

[27] For example, let q be the producer price of the dirty good, and t^{p} the Pigouvian tax based on the standard formula. When the clean good is the untaxed numeraire and labor is taxed, the net wage rate relative to the price of the dirty good is $w(1-\theta)/(q+t^{\mathrm{p}})$. When labor is untaxed, each consumption good faces a tax that raises its price by the factor $\theta/(1-\theta)$, and the dirty good also faces the corrective tax of $t^{\mathrm{p}}/(\mu/\lambda) = t^{\mathrm{p}}/(1-\theta)$, so the net wage relative to the price of the dirty good is $w/[q/(1-\theta)+t^{\mathrm{p}}/(1-\theta)] = w(1-\theta)/(q+t^{\mathrm{p}})$.

into goods and leisure. This result is described by Kaplow (1996), building on previous work of Hylland and Zeckhauser (1979).

Kaplow's observation is that the Samuelson rule for public-goods provision is unaffected by the presence of distortionary taxation when preferences are separable and the government uses a nonlinear income tax. The argument has two pieces. First, following the intuition given above for the proportional-tax case, there will be no change in labor supply, so that all of the expenditures on the public good come through reductions in the untaxed numeraire commodity. Hence, there is no tax wedge at the margin between public and private goods. Second, because of the availability of the nonlinear income tax, the distributional consequences of an increase in public goods spending can be offset, so that distributional weights will also be absent from the decision.

To expand on the reasoning Kaplow provides for his result, we present a detailed proof here. Suppose that households vary with respect to wage rates, w, but that each household's preferences take the form $U(v(c,g), 1-L)$, where c is private good consumption, g is the level of the public good, and L is labor supplied. Public goods are financed using a nonlinear tax $T(wL;g)$ on labor income, where T_1 is the household's marginal tax rate. Consider an experiment in which g is increased, with taxes raised on each individual so that net utility is unchanged. (Continuing to spend and tax in this way will eventually lead to an optimal level of public goods provision, if the government persists to the point that marginal revenue from additional spending is zero.) The claim is that this policy results in no change in labor supply.

The household's initial optimum labor supply decision implies that

$$\frac{\partial U}{\partial L} = U_1 v_1 (w - T_1 w) - U_2 = 0 \tag{5.11}$$

and that Equation (5.11) holds as g changes:

$$v_1 \frac{\mathrm{d}U_1}{\mathrm{d}G} + U_1 \left(v_{11} \frac{\mathrm{d}c}{\mathrm{d}g} + v_{12} \right) w(1 - T_1) - U_1 v_1 w \left(T_{11} w \frac{\mathrm{d}L}{\mathrm{d}g} + T_{12} \right) - U_{21} \frac{\mathrm{d}v}{\mathrm{d}g} - U_{22} \frac{\mathrm{d}L}{\mathrm{d}g} = 0. \tag{5.12}$$

The claim is that Equation (5.12) holds with both U and L constant. Note that if U and L remain constant, so must v, and hence U_1. Thus, the claim implies that

$$v_{11} \frac{\mathrm{d}c}{\mathrm{d}g} + v_{12} = \frac{v_1}{1 - T_1} T_{12} \tag{5.13}$$

or, using $\mathrm{d}v/\mathrm{d}g = v_1\, \mathrm{d}c/\mathrm{d}g + v_2 = 0 \Rightarrow \mathrm{d}c/\mathrm{d}g = -(v_2/v_1)$,

$$\frac{\partial (v_2/v_1)}{\partial c} = \frac{1}{1 - T_1} T_{12}. \tag{5.14}$$

By the assumption that L is fixed, $\mathrm{d}c/\mathrm{d}g = \mathrm{d}T/\mathrm{d}g$ and $\mathrm{d}T/\mathrm{d}g = T_2$. Thus, $v_2/v_1 = T_2$. Moreover, this equality does not hold simply at a particular point, but rather at all

points in the income distribution. That is, the functions $v_2/v_1(c, g)$ and $T_2(wL; g)$ are equal for any value of $c = wL - T(wL; g)$. Thus,

$$\frac{\partial(v_2/v_1)}{\partial c} = T_{21} \left.\frac{\mathrm{d}wL}{\mathrm{d}c}\right|_g = T_{21}\frac{1}{1-T_1}. \tag{5.15}$$

Because $T_{12} = T_{21}$, (5.14) holds, consistent with the initial claim.■

Just as in the case previously considered in Section 5.2, a parallel analysis applies to externalities, with the implication that, under the maintained assumptions regarding preferences and the use of the nonlinear income tax, no adjustment to the standard Pigouvian tax formula is warranted. While these results do depend on two key assumptions, those concerning the separability of individual preferences and the flexibility of the income tax, they are still quite important because they identify the source of deviations from the basic rules of Samuleson and Pigou. As discussed in this Handbook's chapter 25 by Kaplow and Shavell, they also have additional implications regarding the extent to which government policies should be influenced by distributional issues.

6. Optimal taxation and imperfect competition

The analysis to this point concerns the optimal design of tax policies in economies with perfectly competitive industries. Since some economic situations are characterized by imperfect competition, it is useful to consider the implications of differing degrees of market competition for optimal tax design. One of the difficulties of summarizing the implications of imperfect competition for optimal taxation stems from the multiplicity of imperfectly competitive market structures. Nevertheless, it is possible to identify common welfare implications by considering a range of tax instruments and market situations. Our analysis follows closely that of Auerbach and Hines (2001).

6.1. Optimal commodity taxation with Cournot competition

It is useful to start with the behavior of a firm that acts as a Cournot competitor in an industry with a fixed number (n) of firms. The government imposes a specific tax on output at rate t, so firm i's profit is given by

$$Px_i - tx_i - C(x_i), \tag{6.1}$$

in which P is the market price of the firm's output, x_i the quantity it produces, and $C(x_i)$ the cost of producing output level x_i. In this partial-equilibrium setting, it is appropriate to take P to be a univariate function of industry output, denoted X.

The firm's first-order condition for profit maximization is

$$P + x_i \frac{dP}{dX}(1+\theta) - t = C'(x_i), \tag{6.2}$$

in which θ is firm i's conjectural variation, corresponding to $(dX/dx_i - 1)$. Differing market structures correspond to differing values of θ. In a Cournot–Nash setting, in which firm i believes that its quantity decisions do not affect the quantities produced by its competitors, then $\theta = 0$. In a perfectly competitive setting, $\theta = -1$. Various Stackelberg possibilities correspond to values of θ that can differ from these, and indeed, need not lie in the $[-1, 0]$ interval.

It is useful to consider the pricing implications of Equation (6.2). Differentiating both sides of Equation (6.2) with respect to t, taking θ to be unaffected by t, and limiting consideration to symmetric equilibria (so that $x_i = X/n$, $C(x_i) = C(X/n)$, and, since $\frac{dX}{dt} = \frac{dP/dt}{dP/dX}$, it follows that $\frac{dx_i}{dt} = \frac{dP/dt}{n\,dP/dX}$), then

$$\frac{dP}{dt} = \left\{1 + \frac{1+\theta}{n}(1+\eta) - \frac{C''(X/n)}{n\,dP/dX}\right\}^{-1}, \tag{6.3}$$

in which $\eta \equiv \frac{d^2P}{dX^2}\frac{X}{dP/dX}$ is the elasticity of the inverse demand function for X. From Equation (6.3), it is clear that dP/dt can exceed unity, a possibility that is consistent with the firm's second-order condition for profit maximization and with other conditions [discussed by Seade (1980a,b)] that correspond to industry stability.

Equations (6.2) and (6.3) identify the potential welfare impact of taxation in the presence of imperfect competition. From Equation (6.2), the combination of imperfect competition ($\theta > -1$) and a downward-sloping inverse demand function ($dP/dX < 0$) implies that firms choose output levels at which price exceeds marginal cost. Hence there is deadweight loss in the absence of taxation, and, in this simple partial-equilibrium setting, tax policies that stimulate additional output reduce deadweight loss, while those that reduce output make bad situations worse. In some circumstances the imposition of a tax may reduce industry output sufficiently that after-tax profits actually rise.

Tax policy can be used to reduce or eliminate the allocative inefficiency due to imperfect competition, though other policy instruments (such as antitrust enforcement) are also typically available and may be more cost-effective at correcting the problem[28]. Taking alternative remedies to be unavailable, the optimal policy, if the government has access to lump-sum taxation, is to guarantee marginal cost pricing by setting $t = \frac{X}{n}\frac{dP}{dX}(1+\theta)$[29]. Since $dP/dX < 0$, this corrective method entails subsidizing

[28] One possibility, explored by Katz and Rosen (1985), is that tax authorities design corrective policies on the basis of imperfect understanding of the extent of competition in oligopolistic industries.

[29] Such a corrective subsidy was proposed by [Robinson (1933), pp. 163–165], who attributes it to her husband and presents it as an "ingenious but impractical scheme".

the output of the imperfectly competitive industry, so in realistic situations in which tax revenue is obtained through distortionary instruments, it follows that the optimal policy may not fully eliminate the problems due to imperfect competition.

In order to explore this issue further, consider the setup of Section 3.1, in which all commodities are produced at constant cost. There are $N+1$ commodities, of which the first $M+1$, indexed $0,\ldots,M$, are produced by perfectly competitive firms, and the remaining commodities, $M+1,\ldots,N$, are produced in imperfectly competitive markets, each of whose pricing satisfies condition (6.2)[30]. Denoting the (constant) per-unit production cost of commodity i by q_i, it follows that $p_i = q_i + t_i \ \forall\ i = 0, \ldots, M$. As in Section 3, we assume that the tax on the numeraire commodity, good 0, equals 0. Firms in the imperfectly competitive industries generate profits, and someone in the economy receives these profits as income[31]. Taking consumers in the economy to be identical, it follows that the utility of the representative consumer can be represented by

$$V(\boldsymbol{p}, \pi), \tag{6.4}$$

in which $\boldsymbol{p}$ is the vector of $N+1$ commodity prices, and π represents profits earned by the imperfectly competitive firms. Commodity demands are then functions of $(\boldsymbol{p}, \pi)$, but to simplify the calculations that follow, we consider the case in which firms ignore the indirect impact of their pricing decisions on demand through induced changes in profits[32]. In industry $j > M$, the representative firm's first-order condition for profit maximization is

$$p_j - t_j - q_j = -\frac{X_j}{n_j}\frac{(1+\theta_j)}{\partial X_j/\partial p_j}, \tag{6.5}$$

where n_j and θ_j are defined for industry j in the usual way. Thus, the price–cost margin imposed by imperfect competition is

$$m_j = -\frac{X_j(1+\theta_j)}{n_j(\partial X_j/\partial p_j)}$$

in industry j.

[30] We follow much of the literature in assuming that preferences and technology support a unique stable market equilibrium, which, as Roberts and Sonnenschein (1977) note, need not exist in the presence of imperfect competition.

[31] In the competitive context, assuming a zero tax rate on one commodity restricts the government effectively from imposing a tax on pure profits through a uniform tax on all commodities. Here, though, before-tax profits would respond to such uniform taxation, leaving the government's problem unchanged. We show this below, after presenting an expression for equilibrium profits.

[32] This simplifying assumption does not affect the results derived in Equation (6.11) that follows, or their interpretation, but does affect optimal taxes and equilibrium mark-ups.

The optimal taxation problem consists of maximizing (6.4) with respect to the specific taxes $\boldsymbol{t}$ subject to these mark-up conditions, the revenue constraint,

$$\sum_{j=1}^{N} t_j X_j = R, \tag{6.6}$$

and the household's budget constraint,

$$\sum_{j=M+1}^{N} (p_j - t_j - q_j) X_j = \pi. \tag{6.7}$$

Combining the revenue constraint (6.6) and the budget constraint (6.7), we may recast the problem as one of maximizing (6.4) with respect to consumer prices $\boldsymbol{p}$, subject to the constraint

$$\sum_{j=1}^{N} (p_j - q_j) X_j \geqslant R + \pi, \tag{6.8}$$

where profits are given by[33]

$$\pi = -\sum_{j=M+1}^{N} \frac{X_j}{n_j} \frac{(1+\theta_j)}{\partial X_j / \partial p_j} X_j. \tag{6.9}$$

With μ defined as the multiplier of the constraint (6.8), the first-order conditions for this problem are

$$-\lambda X_i + \lambda \frac{\mathrm{d}\pi}{\mathrm{d}p_i} + \mu \left[X_i + \sum_j (p_j - q_j) \frac{\partial X_j}{\partial p_i} + \sum_j (p_j - q_j) \frac{\partial X_j}{\partial y} \frac{\mathrm{d}\pi}{\mathrm{d}p_i} - \frac{\mathrm{d}\pi}{\mathrm{d}p_i} \right] = 0, \quad i = 1, \ldots, N, \tag{6.10}$$

[33] Examination of expression (6.9) clarifies that taxing all goods uniformly would not reduce real profits. Taxing all goods at the same rate would raise prices by a factor λ, so it is necessary to verify that Equation (6.9) continues to hold if profits, π, simultaneously increased by λ (and were therefore unchanged in real terms). Multiplying prices and profits by λ has no effect on X_j, since consumer demands are homogeneous of degree zero in income and prices. But this magnification of prices and income multiplies $\partial X_j / \partial p_j$ by the factor $1/\lambda$, as a unit change in price represents only $1/\lambda$ as large a proportional change as before. Thus, the right-hand side of Equation (6.9) equals its original value, multiplied by λ. As the left-hand side of Equation (6.9) also equals its original value (π) multiplied by λ, the expression still holds.

where, as before, λ is the marginal utility of income. Now defining

$$\alpha^* = \lambda + \mu \sum_{j=1}^{N} t_j^* \frac{\partial X_j}{\partial \pi}$$

to be the "social" marginal utility of income, inclusive of its effect on profits, we may rewrite Equation (6.10) as

$$-\lambda X_i - \mu \left[X_i + \sum_{j=1}^{N} t_j^* \frac{\partial X_j}{\partial p_i} - \left(\frac{\mu - \alpha^*}{\mu} \right) \frac{\mathrm{d}\pi}{\mathrm{d}p_i} \right] = 0 \tag{6.11}$$

in which

$$t_j^* = \begin{cases} t_j & \text{for } j \leqslant M, \\ p_j - q_j & \text{for } j > M, \end{cases}$$

is the total wedge in market j, equal to $t_j + m_j$ in noncompetitive industries. Equation (6.11) is analogous to (5.10), and carries precisely the interpretation offered by Sandmo for the optimal tax conditions in the presence of externalities. Intuitively, the "externality" in the case of imperfect competition is the outcome of the oligopolistic output selection, resulting in the extra mark-up m_j. The definition of t_j^* takes into account the need to correct this pre-existing distortion. Were this the only term in brackets on the left-hand side of Equation (6.11), then it would be optimal fully to correct for the extra distortions in noncompetitive industries and then impose the standard optimal taxes. Presumably, the net result would be an incomplete offset of oligopolistic mark-ups, the optimal tax component normally being positive. The second term in the brackets in Equation (6.11) accounts for the existence of profits, taking the form laid out in expression (3.17) above and explained in that context[34]. In this instance, tax-induced price changes affect the profitability of the imperfectly competitive industry, the difference $(\mu - \alpha^*)$ capturing the welfare effect of increasing industry profits by one unit. To the extent that a higher price of a commodity directly or indirectly augments oligopoly profits, this must be included in computing the price change's overall welfare effect. Doing so has the effect of making the price increase less attractive as a policy tool.

6.2. Specific and ad valorem taxation

In competitive markets the distinction between specific and ad valorem taxation arises only from minor tax enforcement considerations. In imperfectly competitive

[34] Auerbach and Hines (2001) present a longer, different derivation of Equation (6.11) that includes explicit expressions for the terms $\mathrm{d}\pi/\mathrm{d}p_i$. Myles (1989) offers an alternative set of first-order conditions characterizing the set of optimal corrective specific taxes.

markets these two tax instruments are no longer equivalent, since the imposition of an ad valorem tax makes the tax rate per unit of sales a function of a good's price, which is partly under the control of individual firms. As a result, ad valorem and specific taxes that raise equal tax revenue will typically differ in their implications for economic efficiency, ad valorem taxation being associated with much less deadweight loss[35]. Intuitively, ad valorem taxation removes a fraction (equal to the ad valorem tax rate) of a firm's incentive to restrict its output level in order to raise prices.

The welfare superiority of ad valorem taxation is evident in the simple partial-equilibrium setting considered initially above. Now, the government is assumed to have access both to an ad valorem tax and to a specific tax, and tax revenues are assumed costly to obtain (for reasons omitted from the model). In this setting the firm's profits equal

$$(1-\tau)Px_i - tx_i - C(x_i), \tag{6.1'}$$

in which τ is the ad valorem tax rate. Assuming the n-firm outcome to be symmetric, the first-order condition for profit maximization becomes

$$(1-\tau)\left[P + \frac{X}{n}\frac{dP}{dX}(1+\theta)\right] - t = C'\left(\frac{X}{n}\right), \tag{6.2'}$$

and its pricing implications are

$$\frac{dP}{dt} = \left\{(1-\tau)\left[1 + \frac{1+\theta}{n}(1+\eta)\right] - \frac{C''(x/n)}{n\,dP/dX}\right\}^{-1}, \tag{6.12}$$

$$\frac{dP}{d\tau} = \left[P + \frac{X}{n}\frac{dP}{dX}(1+\theta)\right]\frac{dP}{dt}. \tag{6.13}$$

Since a unit change in τ raises more tax revenue than does a unit change in t, it is unsurprising that $\frac{dP}{d\tau} > \frac{dP}{dt}$. Much more revealing is the effect of these tax instruments normalized by dollar of marginal tax revenue. Since total tax revenue is given by $\text{Rev} = \tau PX + tX$, it follows that

$$\frac{d\,\text{Rev}}{dt} = X\left(1 + \tau\frac{dP}{dt}\right) + (t+\tau P)\frac{\partial X}{\partial P}\frac{dP}{dt}, \tag{6.14a}$$

$$\frac{d\,\text{Rev}}{d\tau} = PX\left(1 + \frac{\tau}{P}\frac{dP}{d\tau}\right) + (t+\tau P)\frac{\partial X}{\partial P}\frac{dP}{d\tau}. \tag{6.14b}$$

[35] Suits and Musgrave (1953) provide a classic analysis of this comparison; their treatment is greatly expanded and elaborated by Delipalla and Keen (1992).

In this simple partial-equilibrium model, the change in deadweight loss associated with one of these tax changes is equal to the product of the induced change in X and the difference between marginal cost and price. Consequently,

$$\frac{\mathrm{d}(DWL)/\mathrm{d}t}{\mathrm{d}(DWL)/\mathrm{d}\tau} = \frac{-(\partial X/\partial P)\,(\mathrm{d}P/\mathrm{d}t)}{-(\partial X/\partial P)\,(\mathrm{d}P/\mathrm{d}\tau)}\frac{P - C'\left(\frac{X}{n}\right)}{P - C'\left(\frac{X}{n}\right)} = \frac{(\mathrm{d}P/\mathrm{d}t)}{(\mathrm{d}P/\mathrm{d}\tau)},$$

which, together with Equations (6.14a) and (6.14b), implies that

$$\frac{\frac{\mathrm{d}(DWL)/\mathrm{d}t}{\mathrm{d}(DWL)/\mathrm{d}\tau}}{\frac{\mathrm{d}\,\mathrm{Rev}/\mathrm{d}t}{\mathrm{d}\,\mathrm{Rev}/\mathrm{d}\tau}} = \frac{X\left(\frac{P}{\mathrm{d}P/\mathrm{d}\tau} + \tau\right) + (t + \tau P)\frac{\partial X}{\partial P}}{X\left(\frac{1}{\mathrm{d}P/\mathrm{d}t} + \tau\right) + (t + \tau P)\frac{\partial X}{\partial P}}. \tag{6.15}$$

From Equation (6.13), $\frac{\mathrm{d}P}{\mathrm{d}\tau} < P\frac{\mathrm{d}P}{\mathrm{d}t}$, so if tax revenue is an increasing function of tax rates, then the right-hand side of Equation (6.15) is greater than unity. Hence revenue-equal substitution of ad valorem for specific taxation reduces deadweight loss at any (t, τ) combination[36]. Of course, such substitution works at the expense of firm profitability, and would, if used excessively, drive profits negative and supply presumably to zero. But assuming the firm-profitability constraint not to bind, the optimal tax configuration entails ad valorem rather than specific taxation.

Following the analysis of specific taxes, we seek to maximize the indirect utility function in Equation (6.4) subject to the revenue constraint,

$$\sum_{j=1}^{N} \tau_j p_j X_j \geqslant R, \tag{6.16}$$

the definition of profits,

$$\sum_{j=M+1}^{N} \left(p_j(1 - \tau_j) - q_j\right) X_j = \pi, \tag{6.17}$$

and the characterization of producer behavior in noncompetitive industries,

$$p_j(1 - \tau_i) - q_j = -(1 - \tau_j)\frac{X_j}{n_j}\frac{\left(1 + \theta_j\right)}{\partial X_j/\partial p_j}, \quad j > M. \tag{6.18}$$

[36] Consequently, if the government is able to impose negative specific taxes (specific subsidies), then it can completely eliminate the distortion due to imperfect competition through a judicious combination of ad valorem tax and specific subsidy, as noted by Myles (1996). The effectiveness of this corrective method is limited by any constraints on specific tax rates, such as a restriction that they be nonnegative.

As before, we combine household and government budget constraints to express economy's resource constraint as

$$\sum_{j=1}^{N} (p_j - q_j) X_j \geqslant R + \pi, \tag{6.19}$$

and analyze the problem as one of maximizing (6.4) with respect to $\boldsymbol{p}$, subject to this constraint, where profits are given by

$$\pi = \sum_{j=M+1}^{N} \left[\frac{q_j}{p_j - \phi_j} \right] \phi_j X_j, \quad \text{where} \quad \phi_j = -\frac{X_j}{n_j} \frac{(1+\theta_j)}{\partial X_j / \partial p_j}. \tag{6.20}$$

Note that expression (6.20) differs from (6.9) by the term multiplying $\phi_j X_j$ on the right-hand side of Equation (6.20), which equals $(1 - \tau_j)$. Otherwise, the problem is identical to that for specific taxes, and the first-order conditions (6.11) still hold, for τ_i inserted in place of t_i/p_i. The resulting equilibrium will generally be different, of course, because profits, and hence the terms $\mathrm{d}\pi/\mathrm{d}p_i$, will be different.

Auerbach and Hines (2001) provide some numerical simulations confirming that, in cases for which a noncompetitive industry's tax is positive under specific taxation, it should be higher in the case of ad valorem taxation. They also extend the analysis to the case in which the government is uncertain about the degree of noncompetitive behavior, as represented by the parameter θ. This uncertainty tends to reduce the extent of the desired corrective subsidy, for the subsidy tends to be most effective precisely when it is least needed, i.e., when θ is small.

6.3. *Free entry*

The. standard Cournot model takes as its point of departure an industry with a fixed number of firms. The ability of firms to enter and leave an industry changes the optimal tax problem, and introduces some interesting features of the solution (such as the possibility of welfare-improving positive tax rates even if the government has access to nondistortionary sources of revenue). In spite of these differences, many of the main implications of the preceding analysis, including the welfare superiority of ad valorem to specific taxation, persist in a model with free entry

Consider an industry consisting of identical firms that behave according to (6.2′). In this model, entry and exit are free, but new entrants do not necessarily select output levels that minimize cost, since they behave in a manner that is cognizant of the effect of output on price[37]. The government imposes ad valorem and specific taxes,

[37] New entrants are assumed to exhibit the same oligopolistic behavior (as reflected in θ) as do other firms in the industry; see Mankiw and Whinston (1986) for an analysis of the welfare effects of entry in such a setting.

so the zero-profit condition for industry entry (assuming, for convenience, that it is possible to have fractional numbers of firms) is

$$(1-\tau)\left(\frac{X}{n}\right)P-\left(\frac{X}{n}\right)t-C\left(\frac{X}{n}\right)=0. \tag{6.21}$$

Assuming that the government has access to lump-sum tax instruments, the social total cost (TC) of producing industry output is given simply by its resource cost, or $\text{TC} = nC\left(\frac{X}{n}\right)$. For a small change in a tax instrument, ξ (either an ad valorem or a specific tax), it follows that

$$\frac{\mathrm{d}\text{TC}}{\mathrm{d}\xi}=C'\left(\frac{X}{n}\right)\frac{\mathrm{d}X}{\mathrm{d}\xi}+\left[C\left(\frac{X}{n}\right)-\frac{X}{n}C'\left(\frac{X}{n}\right)\right]\frac{\mathrm{d}n}{\mathrm{d}\xi}. \tag{6.22}$$

The value to consumers for which the tax change is responsible is given by $P\frac{\mathrm{d}X}{\mathrm{d}\xi}$. Consequently, the change in the difference between consumer value and social cost, say Λ, is

$$\frac{\mathrm{d}\Lambda}{\mathrm{d}\xi}=\left[P-C'\left(\frac{X}{n}\right)\right]\frac{\mathrm{d}X}{\mathrm{d}\xi}-\left[\frac{C\left(\frac{X}{n}\right)}{(X/n)}-C'\left(\frac{X}{n}\right)\right]\frac{X}{n}\frac{\mathrm{d}n}{\mathrm{d}\xi}. \tag{6.23}$$

Equation (6.23) succinctly captures the two competing considerations in changing a tax rate that applies to imperfectly competitive industries. The first term is the product of the induced change in output and the difference between price and marginal cost of production for firms in the industry. If the number of firms in the industry were fixed, then this would be the only expression on the right-hand side of Equation (6.23), and it would carry the previous implication that, with the availability of lump-sum tax instruments, efficient taxation consists of equating price and marginal cost. The difficulty, of course, is that it is not the only term on the right-hand side of (6.23). In this model it is necessary to subsidize an industry in order to equate price and marginal cost, and government subsidies encourage inefficient entry of new firms. The welfare effect of tax policy on entry is captured by the second term on the right-hand side of Equation (6.23). This term is the product of the amount of output produced by new entrants and the difference between average and marginal costs for each firm in the industry. Subtracting Equation (6.2′) from (6.21) implies that

$$\frac{C\left(\frac{X}{n}\right)}{X/n}-C'\left(\frac{X}{n}\right)=-(1-\tau)\frac{X}{n}\frac{\mathrm{d}P}{\mathrm{d}X}(1+\theta)>0, \tag{6.24}$$

which simply follows from the fact that price exceeds marginal cost. Hence average cost exceeds marginal cost, and new entry is inefficient, since marginal output is less expensively produced by existing firms than by new entrants[38].

[38] This equilibrium condition requires the production technology to exhibit decreasing average costs over some range of output.

The effect of introducing taxes can be identified by differentiating the identity that $X \equiv nX/n$, which yields

$$\frac{dX}{d\xi} = \left(\frac{X}{n}\right)\frac{dn}{d\xi} + n\frac{d\left(\frac{X}{n}\right)}{d\xi}. \tag{6.25}$$

Together, Equations (6.21), (6.23) and (6.25) imply

$$\frac{d\Lambda}{d\xi} = [P\tau + t]\frac{dX}{d\xi} + \left(\frac{N}{X}\right)\left[nC'\left(\frac{X}{n}\right) - XC'\left(\frac{X}{n}\right)\right]\frac{d\left(\frac{X}{n}\right)}{d\xi}. \tag{6.26}$$

Starting from $t = \tau = 0$, it follows from Equations (6.26) and (6.24) that $d\Lambda/d\xi > 0$ if $d(X/n)/d\xi > 0$, regardless of the effect of taxation on entry and exit. The intuition behind this result is that, while greater output by existing firms promotes efficiency (since price exceeds marginal cost), in the absence of taxation, price equals average cost and there is no welfare impact of marginal entry.

Recall from Equation (6.24) that average cost exceeds marginal cost in equilibrium, and hence is a declining function of a firm's output. Therefore, increases in output per firm will reduce average cost and increase welfare. From the zero-profit condition (6.21), average cost is

$$\text{AC}\left(\frac{X}{n}\right) = \frac{C\left(\frac{X}{n}\right)}{X/n} = (1-\tau)P - t. \tag{6.27}$$

Hence output per firm rises, and therefore welfare rises, in response to the introduction of taxes that reduce the right-hand side of Equation (6.27).

Equation (6.2′) describes the firm's first-order condition for profit maximization. By Equation (6.27), average output per firm (X/n) can be expressed as a decreasing function of $[P(1-\tau) - t]$, while the market demand curve allows us to express total output, X, as a function of P. Appropriately differentiating both sides of Equation (6.2′) with respect to t, evaluating the resulting expression at $\tau = t = 0$, and collecting terms yields

$$\frac{dP}{dt} = \frac{1 + \frac{d(X/n)}{d(P-t)}\frac{dP}{dX}(1+\theta) - C''\left(\frac{X}{n}\right)\frac{d(X/n)}{d(P-t)}}{1 + \frac{d(X/n)}{d(P-t)}\frac{dP}{dX}(1+\theta) - C''\left(\frac{X}{n}\right)\frac{d(X/n)}{d(P-t)}\frac{\eta(1+\theta)}{n}}, \tag{6.28}$$

where η is the elasticity of the inverse demand function, as defined below Equation (6.3). Since the conditions for industry stability imply that both the numerator and the denominator of the expression on the right-hand side of Equation (6.28) are

positive[39], it follows that $(dP/dt - 1)$ has the same sign as $-\eta$. Hence a positive value of η implies that the introduction of a (positive) specific tax increases the market price by less than the amount of the tax, expanding per-firm output and thereby improving welfare[40]. The reason is that the reduced industry output due to a higher tax rate reduces dP/dX, which is a factor in the oligopolistic mark-up by which price is elevated above marginal cost. While the same consideration applies in other settings, the existence of free entry and exit is critical to the welfare result due to the induced attenuation of the effect of taxes on price.

Ad valorem taxation continues to be more attractive than specific taxation in industries with free entry and exit. Starting from $\tau = t = 0$, the introduction of an ad valorem tax reduces the right-hand side of Equation (6.27) if $dP/d\tau < P$. Appropriately differentiating both sides of Equation (6.2′) with respect to τ yields

$$\frac{dP}{d\tau} = P\left\{\frac{1 + \dfrac{d(X/n)}{d[(P(1-\tau)]}\dfrac{dP}{dX}(1+\theta) - C''\left(\dfrac{X}{n}\right)\dfrac{d(X/n)}{d[P(1-\tau)]} + \left(\dfrac{X}{n}\right)\dfrac{dP}{dX}\dfrac{(1+\theta)}{P}}{1 + \dfrac{d(X/n)}{d[(P(1-\tau)]}\dfrac{dP}{dX}(1+\theta) - C''\left(\dfrac{X}{n}\right)\dfrac{d(X/n)}{d[P(1-\tau)]} + \dfrac{\eta(1+\theta)}{n}}\right\}. \tag{6.29}$$

Since the stability conditions imply that the denominator of the right-hand side of Equation (6.29) is positive, it follows that $\left(\frac{dP}{d\tau} - P\right)$ has the same sign as $\left(\frac{dP}{dX}\frac{X}{P} - \eta\right)$. Hence the introduction of an ad valorem tax improves welfare not only if η is positive, but also if η is negative but smaller in absolute value than the elasticity of the inverse demand function. This condition for welfare improvement is weaker than that for the introduction of specific taxes, thereby reflecting the relatively more potent effect of ad valorem taxes in reducing an imperfectly competitive firm's return from restricting output in order to elevate price.

6.4. Differentiated products

Certain types of oligopolistic situations take the form of competition among firms selling products that are imperfect substitutes. Firms take actions that affect product attributes as well as output levels, and these actions are potentially affected by tax policies. Since there are many forms of competition between sellers of differentiated products, it can be difficult to draw general welfare conclusions

[39] Seade (1980a) demonstrates that stability requires $C''(X/n) > (1 + \theta)\,dP/dX$, and since $d(X/n)/d(P - t) < 0$, it follows that the numerator of Equation (6.28) is positive. Seade (1980b) also adopts $\eta + n/(1 + \theta) > 0$ as a stability condition, noting (1980a) that it is a sufficient condition for a firm's marginal revenue to fall when other firms expand output, and that this condition implies that new entry is associated with greater industry output. Together, these stability conditions guarantee that the denominators of Equations (6.28) and (6.29) are positive.

[40] See Besley (1989) and Delipalla and Keen (1992) for additional results and interpretation.

concerning the impact of taxation in such settings; it is, however, possible to identify the major considerations on which the results turn.

Consider an industry of n firms selling products that differ along a univariate quality scale, indexed by v, so that firm i sells products of quality v_i, in which v_i represents a profit-maximizing choice made by the firm. Firm i produces output x_i at quality level v_i, with idiosyncratic costs given by $c_i(x_i, v_i)$. The representative consumer's preferences are then responsible for the inverse demand function $\boldsymbol{p}(\boldsymbol{x}, \boldsymbol{v})$, and the government imposes an ad valorem tax at a uniform rate on all sales in the industry.

Production takes place in two stages. First, firms select values of v_i, taking as fixed the elements of the $\boldsymbol{v}$ vector other than v_i (interesting generalizations are possible by incorporating strategic interaction in the choice of $\boldsymbol{v}$). Second, firms choose output levels x_i contingent on $\boldsymbol{v}$ and taking the output of other firms as fixed. Of course, first-stage choices of $\boldsymbol{v}$ are made in anticipation of induced pricing and output effects in the second stage. Conditional on $\boldsymbol{v}$, firm i's optimal choice of x_i in the second stage must satisfy

$$(1-\tau)\left(p_i(x, v)+\frac{\partial p_i(x, v)}{\partial x_i} x_i\right)=\frac{\partial c_i(x_i, v_i)}{\partial x_i}. \tag{6.30}$$

Denoting the vector of values of x_i that solve (6.30) by $\boldsymbol{x}^*(\boldsymbol{v})$, the first-order condition for the optimal choice of v_i is

$$(1-\tau)\left[\frac{\partial p_i(\boldsymbol{x}^*(\boldsymbol{v}), \boldsymbol{v})}{\partial v_i}+\sum_{j \neq i} \frac{\partial p_i(\boldsymbol{x}^*(\boldsymbol{v}), \boldsymbol{v})}{\partial x_j} \frac{\partial x_j(\boldsymbol{v})}{\partial v_i}\right] x_i=\frac{\partial c_i(x_i, v_i)}{\partial v_i}. \tag{6.31}$$

Oligopolistic situations offer differing interpretations of the context and welfare interpretations of Equations (6.30) and (6.31). From Equation (6.30), it is clear that, conditional on $\boldsymbol{v}$, imperfect competition leads to too little production, in the sense that prices exceed marginal costs. From this observation it is tempting to conclude that (as before) the optimal tax policy is one that subsidizes the output of imperfectly competitive firms. The endogeneity of $\boldsymbol{v}$ has the potential to reverse this reasoning, however, since there is no presumption, from the general form of (6.31), that quality choices are optimal in the absence of taxation.

Quality choice may be suboptimal for many reasons. The first is that firms select quality levels based on their impact on marginal demand and not on the valuation of inframarginal output by the same firm. A second reason is that one firm's return to quality may come at the expense of other firms, and such pecuniary externalities affect welfare in situations in which prices differ from marginal costs. And a third reason is that quality choice in the first stage affects the output decisions of other firms in the second stage, a strategic consideration that creates inefficiencies whenever demand for one commodity is affected by the prices of others.

The examples analyzed in the literature generally share the feature that the introduction of (positive) ad valorem taxation can improve welfare[41]. Equation (6.31) identifies the strategic consideration responsible for this effect, since, if commodities i and j are substitutes in demand ($\partial p_i/\partial x_j < 0$), and strategic substitutes in supply ($\partial x_j/\partial v_i < 0$), then, in the absence of taxation, quality is oversupplied in the sense that $\partial p_i/\partial v_i < \partial c_i/\partial v_i$. Ad valorem taxation typically reduces quality levels, thereby quite possibly improving welfare even though it serves further to distort the output-level choice reflected in Equation (6.30). This implication is very similar to the result (from the previous section) that ad valorem taxation is desirable in a model with free entry and exit, and indeed, these cases share many similarities. Firms described by Equations (6.30) and (6.31) select output levels at which prices exceed marginal costs, but also select quality levels at which marginal costs exceed non-strategic returns. One can think of Equation (6.31) as characterizing excessive "entry" along the quality dimension, and therefore positive ad valorem taxation as being desirable to the extent that it stimulates output per unit of effective quality. Hence, there is potentially a salutary role of taxes in reducing quality, particularly if oligopolistic competition is aggressive in non-price dimensions..

7. Intertemporal taxation

This section considers optimal taxation in intertemporal settings, generally resuming the assumption of perfect competition. Due in part to interest generated by the "consumption tax" advocacy of Fisher and Fisher (1942), Kaldor (1955), and others, one intertemporal issue in particular has received extensive attention: the optimal tax rate on capital income. One of the notable developments of modern optimal tax theory is the finding that, in a simplified second-best setting with identical individuals and in which the government can tax both capital income and labor income, welfare maximization implies zero taxes on capital income in the steady state. This finding reflects, of course, the highly distortionary nature of capital income taxes over long periods of time, but is nevertheless surprising in view of the standard Ramsey intuition that the deadweight loss is zero for the first dollar collected by any tax – and therefore, in the absence of spillovers between markets, all optimal tax rates are strictly positive. Where this intuition fails in the intertemporal context is that it does not account for just how extremely distortionary capital taxation can be even at very low rates of tax – specifically, that low tax rates correspond to distortionary intertemporal tax wedges that grow over time.

[41] See, for example, Kay and Keen (1983) and Cremer and Thisse (1994). Besley and Suzumura (1992) analyze a two-stage game of strategic investment in cost-reducing technology with similar features. Kay and Keen (1991) consider the nature of preferences that determine the effect of taxation on product quality.

The main findings concerning optimal capital taxation are reported by Chamley (1986) and Judd (1985). Subsequent research by Jones, Manuelli and Rossi (1993, 1997), Milesi-Ferretti and Roubini (1998) and others extends its logic to the intertemporal taxation of factors other than capital. In particular, to the extent that wages represent returns to the accumulation of human capital, labor income taxes have capital components and are likewise optimally zero in the steady state. Indeed, the logic of optimal intertemporal taxation is such that there are plausible circumstances in which all taxes may be zero in the steady state. Of course, governments that attempt to implement such optimal taxes would need to amass considerable unspent tax revenue in years prior to the steady state in order to maintain intertemporal budget balance. Before considering these implications, however, it is useful to review the source of the basic intertemporal results concerning capital taxation alone.

7.1. Basic capital income taxation: introduction

The logic of the result that capital is untaxed in the steady state is apparent from working through a simplified version of the Chamley–Judd problem. Consider the case of an economy consisting of identical consumers who maximize the present discounted value of utility over infinite horizons:

$$\sum_{t=0}^{\infty} \beta^t u(C_t, L_t), \tag{7.1}$$

in which β is the one-period discount factor ($\beta = (1+\delta)^{-1}$, δ being an individual's subjective discount rate), taken to be constant for all individuals in all periods. $u(C_t, L_t)$ is a consumer's contemporaneous utility in year t, an increasing function of consumption (C_t) and a decreasing function of labor supplied (L_t).

Consumers have initial wealth of K_0 and earn labor income in period zero equal to $w_0 L_0$, in which w_0 is the after-tax wage rate in period zero. Labor income is received at the start of each period, and consumption also takes place at the start of each period, so any capital income is earned while a period elapses. A consumer therefore dissaves $(C_0 - w_0 L_0)$ in the initial period, and has the lifetime budget constraint

$$\sum_{t=1}^{\infty} (C_t - w_t L_t) \prod_{s=1}^{t} (1 + r_{s-1})^{-1} \leqslant K_0 - (C_0 - w_0 L_0), \tag{7.2}$$

in which r_t is the (after-tax) return earned by capital during period t.

Assuming that the constraint (7.2) is binding (and that the solution entails interior optima), the first-order conditions that characterize the maximum of (7.1) are

$$w_t \frac{\partial u}{\partial C_t} = -\frac{\partial u}{\partial L_t}, \tag{7.3}$$

$$\frac{\partial u}{\partial C_t} = \frac{\partial u}{\partial C_{t+1}} (1 + r_t)\beta. \tag{7.4}$$

Equation (7.4) in turn implies

$$\frac{\partial u}{\partial C_0} = \frac{\partial u}{\partial C_n}\beta^n \prod_{i=0}^{n-1}(1+r_i). \tag{7.5}$$

Combining the budget constraint, (7.2), and the first-order conditions, (7.3) and (7.5), yields

$$\sum_{t=0}^{\infty}\beta^t\left[\frac{\partial u}{\partial C_t}C_t - \frac{\partial u}{\partial L_t}L_t\right] \leqslant K_0\frac{\partial u}{\partial C_0}. \tag{7.6}$$

As the economy consists of identical individuals, we consider the most notationally simple case of one such individual. The period-by-period resource constraint for such an economy is

$$C_t + G_t + K_{t+1} \leqslant F_t(K_t, L_t) + K_t \quad \forall\, t, \tag{7.7}$$

in which G_t is government consumption in period t, and $F(K_t, L_t)$ is the economy's production function. The path of government consumption is taken to be exogenous and (for simplicity) capital is assumed not to depreciate. Inequality (7.7) expresses the idea that the sum of private and public consumption, plus net capital accumulation, cannot exceed the output of the economy.

7.2. *The steady state*

The most straightforward way to evaluate the properties of optimal taxation is to consider the first-order conditions that correspond to maximizing (7.1) subject to (7.6) and (7.7), taking C_t, L_t and K_t to be control variables. [It is noteworthy that Equation (7.7) actually represents a separate constraint for each period.] The first-order condition corresponding to an interior choice of C_t is

$$\beta^t\left\{\frac{\partial u}{\partial C_t} - \lambda\left[\frac{\partial u}{\partial C_t} + C_t\frac{\partial^2 u}{\partial C_t^2} - \frac{\partial^2 u}{\partial L_t\,\partial C_t}L_t\right]\right\} = \mu_t, \tag{7.8}$$

in which λ is the Lagrange multiplier corresponding to the constraint (7.6), and μ_t is the Lagrange multiplier corresponding to condition (7.7) in period t. The first-order condition corresponding to an interior choice of K_t is

$$\mu_t\left(1 + \frac{\partial F}{\partial K_t}\right) = \mu_{t-1}. \tag{7.9}$$

Consider an economy that ultimately settles into a long-run steady state in which economic variables, specifically C_t and L_t, are unchanging. Since the term in braces

on the left-hand side of Equation (7.8) is unchanging in this steady state, it follows that $\mu_t = \beta\mu_{t-1}$. Imposing this equality on Equation (7.9) yields

$$\beta\left(1 + \frac{\partial F}{\partial K_t}\right) = 1. \tag{7.10}$$

Equation (7.4), one of the consumer's first-order conditions, implies that, if $C_t = C_{t+1}$ and $L_t = L_{t+1}$, then $\beta(1 + r_t) = 1$. Consequently, Equation (7.10) implies that $r_t = \partial F/\partial K_t$ in the steady state. Recall that r_t is the after-tax return received by savers during period t. In a competitive market, $\partial F/\partial K_t$ is the pre-tax return to investors. The equality of r_t and $\partial F/\partial K_t$ therefore implies that savings are untaxed.

7.3. Interpreting the solution

The finding that capital income should be untaxed in the steady state contradicts the naïve intuition that, since taxes on labor income distort labor–leisure choices in the steady state, a minor reduction in labor taxes financed by a very small tax on capital income would improve the welfare of the representative individual. Where this intuition fails is that even very low-rate taxes on capital income generate first-order consumption distortions over long horizons. The reason is that a capital-income tax at a very low rate creates a small distortion between consumption in periods t and $(t+1)$, but a large distortion between consumption in period t and consumption in period $(t+n)$, for large n.

It does not by any means follow from the steady-state properties of the optimal program that capital-income taxes are always zero. Indeed, Chamley (1986) offers an example in which consumers have utility functions that are additively separable in consumption and leisure and iso-elastic in consumption, for which the optimal dynamic tax configuration is one in which the government imposes a capital-income tax at a 100 percent rate for an initial period and 0 thereafter[42]. Chamley offers the intuition that high initial rates of capital tax serve to tax away the value of initial capital, thereby acting in part as a lump-sum tax and in part as a very distortionary tax on capital accumulation during the regime of 100 percent tax rates.

This intuitive interpretation of the optimal tax pattern is incomplete, since even if the government possessed an additional tax instrument, permitting it to extract up to 100 percent of the value of initial capital from the private sector, it might still wish to use standard capital-income taxes to raise additional revenue in the short run. The reason is that capital-income taxes in early years distort the choice between present and future consumption, but leave the margins among consumption at different future dates unaffected; nonzero capital-income taxes in later years also distort the pattern of future

[42] Chamley constrains the government not to impose capital income taxes at greater than a 100 percent rate in order to rule out nondistortionary lump-sum initial capital levies as a method of government finance.

consumption. If one thinks of consumption at different dates as separate commodities, then the Ramsey analysis suggests that optimal policy entails equal (revenue-adjusted) marginal distortions to consumption in each period. Because consumption taxes are not included in the government's instrument set, this outcome is approximated by the use of capital-income taxes in early years but not in later years. Analytically, the equations, (7.8) and (7.9), that characterize the optimal path would be formally unchanged even if the government had access to an additional instrument that extracts the value of initial capital. Of course, these conditions would then imply a different tax rate path, but its general feature that capital-income tax rates fall over time would persist, and therefore not reflect the desire to tax the value of initial capital.

The time-varying nature of optimal capital taxation makes such a policy time-inconsistent, in that whatever profile of future taxes that is optimal as of year t would not be optimal as of year $t+1$, and optimizing governments might therefore be tempted not to follow through on previously announced tax plans. Private agents, anticipating such behavior by governments, could not then be expected to respond to announced tax plans in the same way that they would if the government could commit reliably to the taxes that it announces. This is just one of many examples of the time-inconsistency of optimal plans, a feature that takes on special significance in an economy in which private agents hold capital, the value of which governments might find attractive to seize through their tax policies. While there are attempts to identify optimal time-consistent capital-tax policies by somehow constraining government behavior, all such efforts confront the fundamental problem that the mere existence of capital, together with the distortionary nature of income taxation, creates incentives for benevolent governments to behave in a time-inconsistent fashion[43]. The analysis of this section follows the majority of the literature in considering government policies under the assumption that it is possible to make credible commitments.

7.4. Human-capital accumulation and endogenous growth

The model described by Equations (7.1)–(7.10) carries implications for the taxation of labor income, but these are very difficult to characterize succinctly (other than to say that labor-income taxes are positive and unchanging in the steady state). The treatment of labor as a factor of production is somewhat stylized, in that all labor is homogeneous and represents forgone leisure opportunities (with which individuals are endowed).

[43] There is an entirely separate, but relevant, issue that arises concerning the benevolence of governments over time. The optimal tax path is one that accumulates enormous government revenues in the early years in order to finance expenditures in later years (in which capital-income tax rates will be zero). Given the implausibility of actual governments bestowing upon their successors such hard-won budget surpluses in order to finance efficient taxation in the future, it is worth bearing in mind that optimal taxation is a useful ideal if not a reality. In practice, the opposite pattern – in which governments run sizable deficits partly to constrain the fiscal choices of future governments [as in Persson and Svensson 1989)] – is much more common.

The economy described by Equations (7.1)–(7.10) grows via capital accumulation (and shrinks during periods of capital decumulation). As shown by Lucas (1990), Laitner (1995) and others, the qualitative features of optimal taxation are unaffected by introducing exogenous technical progress that generates economic growth and causes the economy to settle into a balanced growth path in the long run. Judd (1999) obtains the similar result that the long-run average optimal capital-income tax rate is likewise zero for economies that do not converge to steady states. Extensions to economies with production subject to stochastic shocks, such as those by Zhu (1992) and Chari, Christiano and Kehoe (1994), produce the result that the optimal tax on capital income is generally very low or zero.

The impact of fiscal policies in settings in which economies grow endogenously is the subject of a closely related literature. There is more than one potential source of endogenous growth, perhaps the most obvious being the accumulation of human capital, along with others that include social increasing returns to scale due to the productivity-enhancing effects of infrastructure and other public goods[44]. These models share the characteristic that the endogeneity of the growth rate arises from some positive externality. As in traditional public-finance analysis, the presence of externalities means that an equilibrium without distortionary taxes will generally not be Pareto-optimal. Thus, optimal tax design must take the presence of such externalities into account, as discussed in Section 5.2.

In some endogenous growth models, the accumulation of human capital generates externalities through intergenerational transmission of acquired skills. However, one may consider the accumulation of human capital and its associated externality separately, and it is useful to do so in understanding the effects on optimal tax results. Human-capital accumulation itself (without any intergenerational transmission of skills) is easily incorporated in the model (7.1)–(7.10), as labor income then represents the return to past forgone consumption and leisure (assuming that both goods and time contribute to the accumulation of human capital), as well as contemporaneous forgone leisure. Since labor-income taxes then effectively tax intertemporal labor/leisure choices in much the same way that capital-income taxes effectively tax intertemporal consumption choices, it is not surprising that the optimal dynamic tax path is one in which labor-income taxes, as well as capital-income taxes, are zero in the steady state [as in Jones, Manuelli and Rossi (1993, 1997) and Milesi-Ferretti and Roubini (1998)].

To show this more formally, consider the case in which consumers have three uses for their time: they can work, for which they receive a wage, they can accumulate human capital, which increases future wages, and they can consume leisure. Denote

[44] See, for example, Lucas (1990), King and Rebelo (1990), Rebelo (1991), Trostel (1993) and Stokey and Rebelo (1995). The sources of endogenous growth analyzed by Eaton (1981) and Hamilton (1987) differ from these in reflecting the saving and portfolio preferences of consumers, and need not entail any productive externalities.

by E_t the amount of time that the consumer devotes to human-capital accumulation in period t. In the simple case in which utility is a function only of consumption and leisure, so that the disutility of time working equals the disutility of devoting the same amount of time to human-capital accumulation, the consumer's maximand becomes

$$\sum_{t=0}^{\infty} \beta^t u(C_t, L_t + E_t). \tag{7.1'}$$

Let H_t denote the consumer's period-t stock of human capital; purely for simplicity assume that human capital does not depreciate. Accumulation of human capital occurs by devoting time and valuable goods and services (e.g., educational resources) to producing additional human capital. Let $M(E,B)$ denote the (time-invariant) human capital production function, in which B represents the value of goods and services devoted to human capital. The accumulation of human capital is therefore constrained by the relationship

$$H_{t+1} \leqslant M(E_t, B_t) + H_t \quad \forall\, t. \tag{7.11}$$

The ability of consumers to allocate some of the economy's output to the accumulation of human capital requires a modification in the economy's resource constraint, as well as a slightly different specification of the production function, so that Equation (7.7) becomes

$$C_t + B_t + G_t + K_{t+1} \leqslant F_t(K_t, L_t, H_t) + K_t \quad \forall\, t. \tag{7.7'}$$

The existence of human capital does not change (7.6), the consumer's intertemporal budget constraint.

The introduction of human capital adds a new state variable (H_t) to the optimal tax problem, as well as two new choice variables (E_t and B_t) and a new constraint (7.11), and requires the modification of the objective function and one of the previous constraints. Once again, the most straightforward way to describe the properties of the optimal solution is to maximize (7.1') subject to (7.6), (7.7') and (7.11), taking C_t, L_t, K_t, B_t and H_t to be control variables. Equations (7.8) and (7.9) continue to hold, and so, therefore, does (7.10) and its implication that the return to saving is untaxed in the steady state.

The first-order condition corresponding to an interior choice of H_t is

$$\mu_t \frac{\partial F}{\partial H_t} + \psi_t = \psi_{t-1}, \tag{7.12}$$

in which ψ_t is the Lagrange multiplier on the constraint (7.11) in period t. The first-order condition corresponding to an interior choice of B_t is

$$\psi_t \frac{\partial M}{\partial B_t} = \mu_t. \tag{7.13}$$

Since Equation (7.8) continues to characterize the optimal solution, it follows that a steady state in which C, L, E and B are unchanging implies that $\mu_t = \beta\mu_{t-1}$. From

Equation (7.13), it then follows that, in the steady state in which $\partial M/\partial B_t$ is unchanging, it must be the case that $\psi_t = \beta\psi_{t-1}$. Together, (7.12), (7.13), and $\psi_t = \beta\psi_{t-1}$ imply

$$1 + \frac{\partial M}{\partial B_t}\frac{\partial F}{\partial H_t} = \frac{1}{\beta}. \tag{7.14}$$

From the steady-state condition $1/\beta = (1 + r)$ it follows that

$$\frac{\partial M}{\partial B_t}\frac{\partial F}{\partial H_t} = r. \tag{7.15}$$

Equation (7.15) characterizes the steady-state economy under optimal taxation, so it is instructive to compare (7.15) to the consumer's first-order conditions. An individual who defers consumption invests either in physical capital or in human capital. Equation (7.4) describes the (interior) first-order condition for investing in physical capital; the analogous first-order condition for investing in human capital is

$$\frac{\partial u}{\partial C_t} = \frac{\partial u}{\partial C_{t+1}}\left(1 + \frac{\partial M}{\partial B_t}\frac{\partial w}{\partial H_t}\right)\beta, \tag{7.16}$$

in which w is the after-tax wage. The term $\partial w/\partial H_t$ in Equation (7.16) therefore equals the single-period after-tax private return from accumulating an additional unit of human capital.

Equations (7.16) and (7.4) together imply that

$$\frac{\partial M}{\partial B_t}\frac{\partial w}{\partial H_t} = r_t,$$

which, together with Equation (7.15), implies that

$$\frac{\partial w}{\partial H_t} = \frac{\partial F}{\partial H_t}. \tag{7.17}$$

The left-hand side of Equation (7.17) is the amount of additional after-tax income received by a worker who accumulates one more unit of human capital; the right-hand side of Equation (7.17) is the marginal product of this additional unit of human capital. Assuming that there are no productivity spillovers, so that the productivity gains from additional human capital are embodied in the effective labor supply of workers who possess the human capital, factor market competition guarantees that the right-hand side of Equation (7.17) equals the effect of human-capital accumulation on pre-tax wages. Since the left-hand side of Equation (7.17) is the effect of human-capital accumulation on after-tax wages, it follows that labor income must be untaxed in the steady state.

Note that this result depends on condition (7.16), which applies only if human-capital accumulation requires inputs of goods – forgone consumption – as well as

leisure. If this is not the case – if human capital is accumulated simply through forgone leisure – then the results that follow will not hold. In particular, the tax on labor income will no longer distort the accumulation of human capital, because the entire cost of investment will be tax deductible. It follows, then, that if goods inputs are deductible, the human-capital decision will remain undistorted by labor-income taxes, in which case there is no requirement that labor-income taxes equal zero in the steady state. As shown by Milesi-Ferretti and Roubini (1998), governments with a sufficient number of tax instruments can effectively decouple the taxation of human-capital accumulation from the taxation of the return to forgone leisure.

The analysis of human-capital accumulation is really a subset of a broader range of issues in which tax instruments are restricted in one way or another. In other settings, Jones, Manuelli and Rossi (1993, 1997) observe that restrictions on the range of tax instruments available to the government, or the presence of public goods in the aggregate production function, change the nature of even steady-state taxation in a way that can make it optimal for the government to impose taxes on capital income. For example, there might be two types of labor in the economy, with properties (such as differing labor-supply elasticities) that would make it optimal to tax the incomes they generate at different rates. If the government is constrained to select a single labor-income tax rate, then the optimal tax rate on capital income might differ from zero in the steady state in order to compensate for the government's inability to tailor its labor-income taxes. Judd (1997) analyzes the implications of restrictions on the ability of the government to control monopolistic and other noncompetitive market behavior, in which case tax policy may function as a different kind of second-best corrective mechanism; his work identifies circumstances under which the optimal tax on capital income may then be negative in the steady state. Coleman (2000) comes to a similar conclusion in a setting in which the government can impose separate consumption and labor-income taxes, and there are restrictions on the range of available tax instruments. Aiyagari (1995) considers the implications of market incompleteness that leaves individuals incapable of diversifying idiosyncratic risks. The resulting demand for precautionary saving leads to a positive optimal tax rate on capital income, even in the steady state.

Correia (1996) notes that many of these considerations stem from the existence of an important productive factor that the government is unable (for some reason) to tax or to subsidize. Depending on the application, this factor might represent inframarginal profits from decreasing returns to scale activity, the returns to monopolistic rents, positive or negative productivity spillovers, labor or capital of specific types, or the value of goods devoted to human-capital accumulation. The effect of such a factor on optimal capital taxation is instructive. Consider the case in which consumers provide an additional productive service, denoted A_t, for which they experience disutility and which the government is unable to tax. The consumer's utility becomes

$$\sum_{t=0}^{\infty} \beta^t u(C_t, L_t, A_t), \tag{7.1''}$$

which the government maximizes subject to the conditions

$$\sum_{t=0}^{\infty} \beta^t \left[\frac{\partial u}{\partial C_t} C_t - \frac{\partial u}{\partial L_t} L_t - \frac{\partial u}{\partial A_t} A_t \right] \leqslant K_0 \frac{\partial u}{\partial C_0} \tag{7.6''}$$

and

$$C_t + G_t + K_{t+1} \leqslant F_t(K_t, L_t, A_t) + K_t. \tag{7.7''}$$

Greater levels of activity A generate pretax returns of $\partial F/\partial A$. The inability of the government to tax the return to A therefore imposes the additional constraint

$$\frac{\partial F}{\partial A_t} \leqslant \frac{\partial U/\partial A_t}{\partial U/\partial C_t}. \tag{7.18}$$

The first-order condition corresponding to an interior choice of C_t is

$$\beta^t \left\{ \frac{\partial u}{\partial C_t} - \lambda \left[\frac{\partial u}{\partial C_t} + C_t \frac{\partial^2 u}{\partial C_t^2} - \frac{\partial^2 u}{\partial L_t \partial C_t} L_t \right] \right\} - \theta_t \frac{\left\{ \frac{\partial^2 U}{\partial A_t \partial C_t} \frac{\partial U}{\partial C_t} - \frac{\partial^2 U}{\partial C_t^2} \frac{\partial U}{\partial A_t} \right\}}{\left(\frac{\partial U}{\partial C_t} \right)^2} = \mu_t, \tag{7.19}$$

in which θ_t is the Lagrange multiplier corresponding to the constraint (7.18). The first-order condition corresponding to an interior choice of K_t is

$$\mu_t \left(1 + \frac{\partial F}{\partial K_t} + \theta_t \frac{\partial^2 F}{\partial A_t \partial K_t} \right) = \mu_{t-1}. \tag{7.20}$$

Taking the Lagrange multiplier θ_t to grow at rate β in the steady state, these conditions together imply that, in the steady state,

$$r_t = \frac{\partial F}{\partial K_t} + \theta_t \frac{\partial^2 F}{\partial A_t \partial K_t}. \tag{7.21}$$

Equation (7.21) is inconsistent with zero capital taxation whenever two conditions hold simultaneously: that constraint (7.18) binds, and that changes in K affect the marginal productivity of A.

In the case of ordinary human-capital accumulation, the government does not seek to tax A (which can be interpreted as past labor effort used to accumulate human capital), so $\theta_t = 0$ and physical capital is untaxed as well. In the case of economies with public goods or other types of productive externalities, or those in which heterogeneous inputs must receive identical tax treatment, a government that cannot use corrective taxation to induce efficient decentralized behavior will change its other taxes to accommodate

the missing market[45]. As a result, steady-state tax rates on capital will be greater than, equal to, or less than zero according to the nature of the externality (positive or negative) and the complementarity or substitutability of the untaxed factor with capital – a standard implication along the lines of Corlett and Hague (1953) in a static setting.

7.5. Results from life-cycle models

Though undoubtedly a powerful and illuminating result, the convergence of the optimal capital-income tax to zero rests on the implausible assumption that agents live forever or behave in an equivalent manner with respect to their heirs. Without infinite lifetimes, no such result holds, although intuition suggests that long but finite lifetimes still would place strong bounds on the size of the optimal capital-income tax. However, with finite lifetimes also comes the complication of heterogeneity with respect to age cohort, which tax-policy optimization must take into account. Thus, there is more to learn from consideration of finite-lifetime, overlapping generation (OG) models than that capital-income taxes should be low, if not zero, in the long run.

The Diamond (1965) model, in which each generation lives for two periods, consuming in both and working in the first, provided the basis for the initial research on optimal taxation in OG models. In this model, without bequests, the lifetime budget constraint for the representative household born in period t may be written

$$C_t^1 + \left(\frac{1}{1+r_{t+1}}\right) C_{t+1}^2 = w_t L_t, \tag{7.22}$$

where C^1 is consumption when young, C^2 is consumption when old, L is labor supply when young, and subscripts indicate periods in which activity occurs.

As is clear from this expression, endowing the government with two instruments, proportional taxes on labor income (which affects w) and capital income (which affects r), is equivalent to allowing the government to tax first- and second-period consumption, at possibly different rates. A zero tax on capital income – a labor-income tax – would result in uniform taxation of consumption in the two periods.

Using this model, papers by Diamond (1973), Pestieau (1974), Auerbach (1979) and Atkinson and Sandmo (1980) characterized optimal steady-state taxes under different assumptions about instruments available to the government. Two general

[45] Auerbach (1979) offers a similar analysis of the optimal taxation of heterogeneous capital goods in the presence of other constraints. Coleman's (2000) analysis of optimal consumption and labor-income taxes takes the path of future government spending to be fixed in nominal terms, which implies that, in the steady state, the combination of a consumption tax and a labor subsidy relaxes the government's revenue requirement by reducing real government spending. Coleman finds that, if the labor-income tax is constrained to be non-negative, then the optimal steady-state labor-income tax rate is zero and the tax on income from capital (which is a substitute for labor) is negative.

results from this literature are that (1) with government debt available to redistribute resources across generations, the marginal product of capital should converge to the intertemporal discount rate embodied in the government's social welfare function; and (2) in this equilibrium, optimal taxes on labor and capital facing individual cohorts should follow the standard three-good analysis of static optimal tax theory, with a zero tax on capital income being optimal only for a certain class of preferences. Result (1) confirms that Cass's (1965) "modified Golden rule" result holds even in the presence of distortionary taxation. It is analogous to the Chamley–Judd result discussed above. However, as result (2) confirms, this does not imply that capital-income taxes converge to zero. The marginal product of capital is being equated to the *government's* discount rate (for comparing the consumption of different cohorts at different points in time), not the discount rate used by individual households in comparing their own first- and second-period consumption.

These results, like those derived for the infinite-lifetime case, tell us little about the nature of optimal tax schedules in transition; nor are they useful in determining how the long-run optimum might differ if the government faced constraints on its short-run policy. For example, if the optimal path for capital-income taxes were one of high taxes declining to zero (as in Chamley's analysis), but the government's decision whether or not to abolish capital-income taxes had to be made on a once-and-for-all basis, would it still improve economic efficiency to abolish capital-income taxes? As transition constraints are a major concern of actual tax policy decisions, understanding the linkage between transition and long-run policy is important.

Analyzing the efficiency (and incidence) effects of tax policies in transition has been a major objective of the literature utilizing dynamic computable general equilibrium (CGE) models based on more realistic characterizations of life-cycle behavior. Auerbach, Kotlikoff and Skinner (1983) and Auerbach and Kotlikoff (1987) developed a 55-generation OG model with endogenous labor supply and retirement, in which agents alive during the transition from one steady state to another have perfect foresight about future factor prices and tax rates. Their central simulations consider the impact of switching immediately from a uniform tax on labor and capital income to a tax on labor income or a consumption tax. While such taxes appear equivalent in terms of the lifetime budget constraint represented in Equation (7.22), as well as in the 55-period version of this budget constraint, they are *not* the same with respect to transition generations, who begin the transition with previously accumulated life-cycle wealth. For these transition generations, a consumption tax is equivalent to a tax on labor income *plus* a tax on existing wealth – a capital levy. This can be seen by considering an amended version of Equation (7.22) that has some measure of existing assets, A_t, on the right-hand side. Thus, the transition to a consumption tax is more attractive than a transition to a labor-income tax from the standpoint of economic efficiency.

Determining the efficiency differences between these two reforms is complicated by the fact that the reforms also have different intergenerational incidence, the consumption tax harming initial generations at the expense of future generations,

the labor-income tax doing the reverse. As a result, the steady-state welfare gain overstates the efficiency gain in the case of the consumption tax, for it reflects not only efficiency gains but also transfers from transition generations. By the same logic, the steady-state welfare gain understates the efficiency gain in the case of the labor-income tax. To separate incidence from efficiency effects, the authors construct a hypothetical "lump-sum redistribution authority" that makes balanced-budget lump-sum taxes and transfers among generations to ensure that all transition generations are kept at the pre-reform utility level and all post-transition generations enjoy an equal increase in lifetime utility, an increase that can be viewed as a measure of the policy's efficiency gain (or loss, if negative). With this adjustment, and for base-case parameter assumptions, the transition to a consumption tax is predicted to increase economic efficiency, while the transition to a labor-income tax would reduce economic efficiency.

The key lesson of these simulations is that tax systems that appear to be equivalent from the perspective of a representative individual may differ significantly in an economy with different age cohorts. A corollary is that adopting a consumption tax but simultaneously providing transition relief for those harmed by the tax in transition will offset not only adverse distributional effects, but also the efficiency benefits of the capital levy. Auerbach (1996) illustrates this result in an analysis of a range of consumption-type tax reform proposals that vary in the extent to which they provide transition relief. The putative efficiency advantage of the consumption tax relies, of course, on the ability of the government to use the implicit capital levy "just once" and raises the question of dynamic inconsistency discussed above.

Just as it is possible to extend the representative-agent, infinite-horizon model to include human-capital accumulation, this has been one direction in which dynamic CGE models have been extended in recent years, most notably by Heckman, Lochner and Taber (1998).

8. Conclusions

The analysis of excess burden and optimal taxation is one of the oldest subjects in applied economics, yet research continues to offer important new insights that build on the original work of Dupuit, Jenkin, Marshall, Pigou, Ramsey, Hotelling, and others. Fundamentally, it remains true that departures from marginal cost pricing are associated with excess burden, that the magnitude of excess burden is roughly proportional to the square of any such departure, and that efficient tax systems are ones that minimize excess burden subject to achieving other objectives. The contribution of modern analysis is to identify new and important reasons for prices and marginal costs to differ, to assess their practical magnitudes, and to consider their implications for taxation.

One of the major developments of the last fifty years is the widespread application of rigorous empirical methods to analyze the efficiency of the tax system. Empirical

work not only assists the formation and analysis of economic policy, but also plays a critical role in distinguishing important from less important theoretical considerations, thereby contributing to further theoretical development. Properly executed, empirical analysis is not only consistent with the welfare theory that underlies normative public finance, but also takes the theory further by testing its implications and offering reliable measurement of parameters that are critical to the assessment of tax systems.

Recognition of the importance of population heterogeneity and of the potential complications of evaluating policy reforms with pre-existing distortions has motivated much of the recent normative work in public finance. The new learning serves generally to highlight the value of Ramsey's insights by demonstrating their application to a variety of settings, including those with population heterogeneity and a wide range of available tax instruments. Mirrlees differs from Ramsey in focussing on the role of informational asymmetries between governments and taxpayers as a determinant of the shape of optimal tax schedules; nevertheless, Ramsey-like conditions characterize optimal tax policy even in this setting.

The efficiency of the tax system is a topic of enduring importance and continuing investigation. Economic analysis has much to offer on the topic of efficiency, and indeed, is occasionally criticized for offering too much. The other chapters in this Handbook offer what is perhaps an illustration of this proposition by examining both positive and normative aspects of taxation in a wide variety of settings.

References

Aiyagari, S.R. (1995), "Optimal capital income taxation with incomplete markets, borrowing constraints, and constant discounting", Journal of Political Economy 106:1158–1175.

Atkinson, A.B., and A. Sandmo (1980), "Welfare implications of the taxation of savings", Economic Journal 90:529–549.

Atkinson, A.B., and N.H. Stern (1974), "Pigou, taxation and public goods", Review of Economic Studies 41:119–128.

Atkinson, A.B., and J.E. Stiglitz (1972), "The structure of indirect taxation and economic efficiency", Journal of Public Economics 1:97–119.

Atkinson, A.B., and J.E. Stiglitz (1976), "The design of tax structure: direct versus indirect taxation", Journal of Public Economics 6:55–75.

Atkinson, A.B., and J.E. Stiglitz (1980), Lectures in Public Economics (McGraw-Hill, New York).

Auerbach, A.J. (1979), "The optimal taxation of heterogeneous capital", Quarterly Journal of Economics 93:589–612.

Auerbach, A.J. (1985), "The theory of excess burden and optimal taxation", in: Alan J. Auerbach and Martin Feldstein, eds., Handbook of Public Economics, Vol. 1 (North-Holland, Amsterdam) pp. 61–127.

Auerbach, A.J. (1996), "Tax reform, capital allocation, efficiency and growth", in: Henry J. Aaron and William G. Gale, eds., Economic Effects of Fundamental Tax Reform (Brookings, Washington, D.C.) pp. 29–81.

Auerbach, A.J., and J.R. Hines Jr (2001), "Perfect taxation with imperfect competition", Working Paper 8138 (National Bureau of Economic Research).

Auerbach, A.J., and L.J. Kotlikoff (1987), Dynamic Fiscal Policy (Cambridge University Press, Cambridge, UK).

Auerbach, A.J., L.J. Kotlikoff and J. Skinner (1983), "The efficiency gains from dynamic tax reform", International Economic Review 24:81–100.
Auten, G., and R. Carroll (1999), "The effect of income taxes on household income", Review of Economics and Statistics 81:681–693.
Ballard, C.L., and D. Fullerton (1992), "Distortionary taxes and the provision of public goods", Journal of Economic Perspectives 6:117–131.
Besley, T. (1989), "Commodity taxation and imperfect competition: a note on the effects of entry", Journal of Public Economics 40:359–366.
Besley, T., and K. Suzumura (1992), "Taxation and welfare in an oligopoly with strategic commitment", International Economic Review 33:413–431.
Bradford, D. (1986), Untangling the Income Tax (Harvard University Press, Cambridge, MA).
Browning, E.K. (1975), "Labor supply distortions of social security", Southern Economic Journal 42:243–252.
Cass, D. (1965), "Optimum growth in an aggregative model of capital accumulation", Review of Economic Studies 32:233–240.
Chamley, C. (1986), "Optimal taxation of capital income in general equilibrium with infinite lives", Econometrica 54:607–622.
Chari, V.V., L.J. Christiano and P.J. Kehoe (1994), "Optimal fiscal policy in a business cycle model", Journal of Political Economy 102:617–652.
Coleman II, W.J. (2000), "Welfare and optimum dynamic taxation of consumption and income", Journal of Public Economics 76:1–39.
Corlett, W.J., and D.C. Hague (1953), "Complementarity and the excess burden of taxation", Review of Economics Studies 10:295–337.
Correia, I.H. (1996), "Should capital income be taxed in the steady state?" Journal of Public Economics 60:147–151.
Cremer, H., and J.-F. Thisse (1994), "Commodity taxation in a differentiated oligopoly", International Economic Review 35:613–633.
Dahan, M., and M. Strawczynski (2000), "Optimal income taxation: an example with a u-shaped pattern of optimal marginal tax rates: comment", American Economic Review 90:681–686.
Deaton, A. (1979), "Optimally uniform commodity taxes", Economic Letters 2:357–361.
Deaton, A., and J. Muellbauer (1980), "An almost ideal demand system", American Economic Review 70:312–326.
Delipalla, S., and M. Keen (1992), "The comparison between ad valorem and specific taxation under imperfect competition", Journal of Public Economics 49:351–366.
Diamond, P.A. (1965), "National debt in a neoclassical growth model", American Economic Review 55:1126–1150.
Diamond, P.A. (1973), "Taxation and public production in a growth setting", in: James A. Mirrlees and Nicholas H. Stern, eds., Models of Economic Growth (Macmillan, New York) pp. 215–235.
Diamond, P.A. (1975), "A many-person Ramsey rule", Journal of Public Economics 4:335–342.
Diamond, P.A. (1998), "Optimal income taxation: an example with a u-shaped pattern of optimal marginal tax rates", American Economic Review 88:83–95.
Diamond, P.A., and J.A. Mirrlees (1971a), "Optimal taxation and public production I: production efficiency", American Economic Review 61:8–27.
Diamond, P.A., and J.A. Mirrlees (1971b), "Optimal taxation and public production II: tax rules", American Economic Review 61:261–278.
Diamond, P.A., L.J. Helms and J.A. Mirrlees (1980), "Optimal taxation in a stochastic economy: a Cobb–Douglas example", Journal of Public Economics 14:1–29.
Dixit, A.K., and A. Sandmo (1977), "Some simplified formulae for optimal income taxation", Scandinavian Journal of Economics 79:417–423.
Dupuit, A.J.É.J. (1844), "De la mesure de l'utilité des travaux publics", Annales des Ponts et Chaussées, 2nd series, 8; translated by R.H. Barback, 1952, "On the measurement of the utility of public works",

International Economic Papers 2:83–110; reprinted, 1969, in: Kenneth J. Arrow and Tibor Scitovsky, eds., Readings in Welfare Economics (Richard D. Irwin, Homewood, IL) pp. 255–283.
Eaton, J. (1981), "Fiscal policy, inflation and the accumulation of risky capital", Review of Economic Studies 48:435–445.
Eaton, J., and H.S. Rosen (1980), "Labor supply, uncertainty, and efficient taxation", Journal of Public Economics 14:365–374.
Feldstein, M. (1978), "The welfare cost of capital income taxation", Journal of Political Economy 86:S29–S51.
Feldstein, M. (1995), "The effect of marginal tax rates on taxable income: a panel study of the 1986 Tax Reform Act", Journal of Political Economy 103:551–572.
Feldstein, M. (1999), "Tax avoidance and the deadweight loss of the income tax", Review of Economics and Statistics 81:674–680.
Fisher, I., and H.W. Fisher (1942), Constructive Income Taxation, A Proposal for Reform (Harper, New York).
Gallant, R.A. (1981), "On the bias in flexible functional forms and an essentially unbiased form: the Fourier functional form", Journal of Econometrics 15:211–245.
Goolsbee, A. (2000), "What happens when you tax the rich? Evidence from executive compensation", Journal of Political Economy 108:352–378.
Håkonsen, L. (1998), "An investigation into alternative representations of the marginal cost of public funds", International Tax and Public Finance 5:329–343.
Hall, R.E., and A. Rabushka (1995), The Flat Tax, 2nd Edition (Hoover Institution Press, Stanford).
Hamilton, J.H. (1987), "Taxation, savings, and portfolio choice in a continuous time model", Public Finance 42:264–282.
Harberger, A.C. (1964a), "The measurement of waste", American Economic Review 54:58–76.
Harberger, A.C. (1964b), "Taxation, resource allocation, and welfare", in: John F. Due, ed., The Role of Direct and Indirect Taxes in the Federal Revenue System (Princeton University Press, Princeton, NJ) pp. 25–70.
Harberger, A.C. (1966), "Efficiency effects of taxes on income from capital", in: Marian Krzyzaniak, ed., Effects of Corporation Income Tax (Wayne State University Press, Detroit) pp. 107–117.
Harberger, A.C. (1971), "Three basic postulates for applied welfare economics: an interpretive essay", Journal of Economic Literature 9:785–797.
Harberger, A.C. (1972), "The opportunity costs of public investment financed by borrowing", in: R. Layard, ed., Cost–Benefit Analysis (Penguin, Harmondsworth, UK) pp. 303–310.
Hausman, J.A. (1981a), "Labor supply", in: Henry J. Aaron and Joseph A. Pechman, eds., How Taxes Affect Economic Behavior (Brookings, Washington) pp. 27–72.
Hausman, J.A. (1981b), "Exact consumer's surplus and deadweight loss", American Economic Review 71:662–676.
Hausman, J.A., and W. Newey (1995), "Nonparametric estimation of exact consumer surplus and deadweight loss", Econometrica 63:1445–1476.
Heckman, J.J., L. Lochner and C. Taber (1998), "Tax policy and human capital formation", American Economic Review 88:293-297.
Hines Jr, J.R. (1999), "Three sides of Harberger triangles", Journal of Economic Perspectives 13:167–188.
Hylland, A., and R.J. Zeckhauser (1979), "Distributional objectives should affect taxes but not program choice or design", Scandinavian Journal of Economics 81:264–284.
Jenkin, H.C.F. (1871/72), "On the principles which regulate the incidence of taxes", in: Proceedings of the Royal Society of Edinburgh, pp. 618–631; reprinted, 1887, in: Sidney Colvin and J.A. Ewing, eds., Papers, Literary, Scientific, etc., by the late Fleeming Jenkin, Vol. 2 (Longmans Green, London) pp. 107–121.
Jones, L.E., R.E. Manuelli and P.E. Rossi (1993), "Optimal taxation in models of endogenous growth", Journal of Political Economy 101:485–517.

Jones, L.E., R.E. Manuelli and P.E. Rossi (1997), "On the optimal taxation of capital income", Journal of Economic Theory 73:93–117.
Jorgenson, D.W., L.J. Lau and T.M. Stoker (1982), "The transcendental logarithmic model of aggregate consumer behavior", in: Robert L. Basmann and G. Rhodes, eds., Advances in Econometrics, Vol. 1 (JAI Press, Greenwich, CT) pp. 97–238.
Judd, K.L. (1985), "Redistributive taxation in a simple perfect foresight model", Journal of Public Economics 28:59–83.
Judd, K.L. (1997), "The optimal tax rate for capital income is negative", Working Paper 6004 (National Bureau of Economic Research).
Judd, K.L. (1999), "Optimal taxation and spending in general competitive growth models", Journal of Public Economics 71:1–27.
Kaldor, N. (1955), An Expenditure Tax (Allen and Unwin, London).
Kanbur, R., and M. Tuomala (1994), "Inherent inequality and the optimal graduation of marginal tax rates", Scandinavian Journal of Economics 96:275–282.
Kaplow, L. (1996), "The optimal supply of public goods and the distortionary cost of taxation", National Tax Journal 49:513–533.
Katz, M.L., and H.S. Rosen (1985), "Tax analysis in an oligopoly model", Public Finance Quarterly 13:3–20.
Kay, J.A. (1980), "The deadweight loss from a tax system", Journal of Public Economics 13:111–120.
Kay, J.A., and M. Keen (1983), "How should commodities be taxed? Market structure, product heterogeneity and the optimal structure of commodity taxes", European Economic Review 23: 339–358.
Kay, J.A., and M. Keen (1991), "Product quality under specific and ad valorem taxation", Public Finance Quarterly 19:238–246.
King, M.A. (1983), "Welfare analysis of tax reforms using household data", Journal of Public Economics 21:183–214.
King, R.G., and S. Rebelo (1990), "Public policy and economic growth: developing neoclassical implications", Journal of Political Economy 98:S126–S150.
Laitner, J. (1995), "Quantitative evaluations of efficient tax policies for Lucas' supply side models", Oxford Economic Papers 47:471–492.
Lindsey, L.B. (1987), "Individual taxpayer response to tax cuts, 1982–1984, with implications for the revenue maximizing tax rate", Journal of Public Economics 33:173–206.
Lucas Jr, R.E. (1990), "Supply-side economics: an analytical review", Oxford Economic Papers 42: 293–317.
Mankiw, N.G., and M.D. Whinston (1986), "Free entry and social inefficiency", Rand Journal of Economics 17:48–58.
Milesi-Ferretti, G.-M., and N. Roubini (1998), "On the taxation of human and physical capital in models of endogenous growth", Journal of Public Economics 70:237–254.
Mirrlees, J.A. (1971), "An exploration in the theory of optimum income taxation", Review of Economics Studies 38:175–208.
Mirrlees, J.A. (1976), "Optimal tax theory: a synthesis", Journal of Public Economics 6:327–358.
Mirrlees, J.A. (1986), "The theory of optimal taxation", in: Kenneth J. Arrow and Michael D. Intriligator, eds., Handbook of Mathematical Economics, Vol. 3 (North-Holland, Amsterdam) pp. 1197–1249.
Moffitt, R., and M. Wilhelm (2000), "Taxation and the labor supply decisions of the affluent", in: Joel B. Slemrod, ed., Does Atlas Shrug? The Economic Consequences of Taxing the Rich (Harvard University Press, Cambridge, MA) pp. 193–234.
Mohring, H. (1971), "Alternative welfare gain and loss measures", Western Economic Journal 9:349–368.
Myles, G.D. (1989), "Ramsey tax rules for economies with imperfect competition", Journal of Public Economics 38:95–115.
Myles, G.D. (1996), "Imperfect competition and the optimal combination of ad valorem and specific taxation", International Tax and Public Finance 3:29–44.

Persson, T., and L.E.O. Svensson (1989), "Why a stubborn conservative would run a deficit: policy with time-inconsistent preferences", Quarterly Journal of Economics 104:325–345.
Pestieau, P.M. (1974), "Optimal taxation and discount rate for public investment in a growth setting", Journal of Public Economics 3:217–235.
Phelps, E.S. (1973), "The taxation of wage income for economic justice", Quarterly Journal of Economics 87:331–354.
Pigou, A.C. (1947), A Study in Public Finance, 3rd Edition (Macmillan, London).
Ramsey, F.P. (1927), "A contribution to the theory of taxation", Economic Journal 37:47–61.
Rebelo, S. (1991), "Long-run policy analysis and long-run growth", Journal of Political Economy 99:500–521.
Roberts, J., and H. Sonnenschein (1977), "On the foundations of the theory of monopolistic competition", Econometrica 45:101–113.
Robinson, J. (1933), The Economics of Imperfect Competition (Macmillan, London).
Rosen, H.S. (1978), "The measurement of excess burden with explicit utility functions", Journal of Political Economy 86:S121–S135.
Sadka, E. (1976), "On income distribution, incentive effects and optimal income taxation", Review of Economic Studies 43:261–268.
Saez, E. (2000a), "Using elasticities to derive optimal income tax rates", Working Paper 7628 (National Bureau of Economic Research).
Saez, E. (2000b), "Optimal income transfer programs: intensive versus extensive labor supply responses", Working Paper 7708 (National Bureau of Economic Research).
Samuelson, P.A. (1951), "Theory of optimal taxation", Unpublished memorandum (U.S. Treasury); published in 1986, Journal of Public Economics 30:137–143.
Samuelson, P.A. (1954), "The pure theory of public expenditure", Review of Economics and Statistics 36:387–389.
Sandmo, A. (1975), "Optimal taxation in the presence of externalities", Swedish Journal of Economics 77:86–98.
Sandmo, A. (1985), "The effects of taxation on savings and risk taking", in: Alan J. Auerbach and Martin Feldstein, eds., Handbook of Public Economics, Vol. 1 (North-Holland, Amsterdam) pp. 265–311.
Sandmo, A. (1998), "Redistribution and the marginal cost of public funds", Journal Public Economics 70:365–382.
Schöb, R. (1997), "Environmental taxes and pre-existing distortions: the normalization trap", International Tax and Public Finance 4:167–176.
Seade, J. (1977), "On the shape of optimal tax schedules", Journal of Public Economics 7:203–237.
Seade, J. (1980a), "On the effects of entry", Econometrica 48:479–489.
Seade, J. (1980b), "The stability of Cournot revisited", Journal of Economic Theory 23:15–26.
Shoven, J.B. (1976), "The incidence and efficiency effects of taxes on income from capital", Journal of Political Economy 84:1261–1283.
Slemrod, J., and S. Yitzhaki (2001), "Integrating expenditure and tax decisions: the marginal cost of funds and the marginal benefit of projects", National Tax Journal 54:189–201.
Slemrod, J., S. Yitzhaki, J. Mayshar and M. Lundholm (1994), "The optimal two-bracket linear income tax", Journal of Public Economics 53:269–290.
Slesnick, D.T. (1998), "Empirical approaches to the measurement of welfare", Journal of Economic Literature 36:2108–2165.
Stern, N.H. (1976), "On the specification of models of optimum income taxation", Journal of Public Economics 6:123–162.
Stern, N.H. (1982), "Optimum taxation with errors in administration", Journal of Public Economics 17:181–211.
Stiglitz, J.E. (1982), "Self-selection and Pareto efficient taxation", Journal of Public Economics 17: 213–240.

Stiglitz, J.E. (1987), "Pareto efficient and optimal taxation and the new new welfare economics", in: Alan J. Auerbach and Martin Feldstein, eds., Handbook of Public Economics, Vol. 2 (North-Holland, Amsterdam) pp. 991–1042.

Stiglitz, J.E., and P.S. Dasgupta (1971), "Differential taxation, public goods and economic efficiency", Review of Economic Studies 38:151–174.

Stokey, N.L., and S. Rebelo (1995), "Growth effects of flat-rate taxes", Journal of Political Economy 103:519–550.

Suits, D.B., and R.A. Musgrave (1953), "Ad valorem and unit taxes compared", Quarterly Journal of Economics 67:598–604.

Trostel, P.A. (1993), "The effect of taxation on human capital", Journal of Political Economy 101: 327–350.

Varian, H.R. (1980), "Redistributive taxation as social insurance", Journal of Public Economics 14:49–68.

Vartia, Y. (1983), "Efficient methods of measuring welfare change and compensated income in terms of ordinary demand functions", Econometrica 51:79–98.

Zhu, X. (1992), "Optimal fiscal policy in a stochastic growth model", Journal of Economic Theory 58:250–289.

Chapter 22

TAX AVOIDANCE, EVASION, AND ADMINISTRATION*

JOEL SLEMROD

The University of Michigan

SHLOMO YITZHAKI

The Hebrew University of Jerusalem

Contents

* We are grateful to Wojciech Kopczuk for valuable research assistance. Helpful comments on an earlier draft were received from Jim Alm, Alan Auerbach, Len Burman, Brian Erard, Martin Feldstein, Firouz Gahvari, Roger Gordon, Jim Hines, Jonathan Kesselman, Louis Kaplow, Jim Poterba, Agnar Sandmo, Dan Shaviro, Eric Toder, Gideon Yaniv, the participants at the NYU Law School Colloquium on Tax Policy and Public Finance and the handbook conference held at the University of California at Berkeley.

Handbook of Public Economics, Volume 3, Edited by A.J. Auerbach and M. Feldstein

Abstract

Tax avoidance and evasion are pervasive in all countries, and tax structures are undoubtedly skewed by this reality. Standard models of taxation and their conclusions must reflect these realities.

This paper first presents theoretical models that integrate avoidance and evasion into the overall decision problem faced by individuals. Early models of this area focused on tax evasion, modeled as a gamble against the enforcement capability of the state. More recently, the literature has examined more general models of the technology of avoidance, with the additional risk bearing caused by tax evasion either being a special case of this technology or one aspect of the cost of changing behavior to reduce tax liability. If the cost of evasion and avoidance depends on other aspects of behavior, the choice of consumption basket and avoidance become intertwined. The paper then relates the behavior predicted by the model to what is known empirically about the extent of evasion and avoidance, and how it responds to tax enforcement policy.

The paper then turns to normative analysis, and discusses how avoidance and evasion affect the analysis of vertical and horizontal equity as well as efficiency costs; a taxonomy of efficiency costs is presented. Acknowledging the variety of behavioral responses to taxation changes the answers to traditional subjects of inquiry, such as incidence, optimal progressivity, and the optimal mix between income and consumption taxes. It also raises a whole new set of policy questions, such as the appropriate level of resources to devote to administration and enforcement, and how those resources should be deployed. Because there are a variety of policy instruments that can affect the magnitude and nature of avoidance and evasion response, the elasticity of behavioral response is itself a policy instrument, to be chosen optimally.

The paper reviews what is known about these issues, and introduces a general theory of optimal tax systems, in which tax rates and bases are chosen simultaneously with the administrative and enforcement regimes. We argue that the concept of the marginal efficiency cost of funds is a useful way to summarize the normative issues that arise, and expand the concept to include administrative costs, avoidance, and evasion.

Keywords

taxation, evasion, enforcement, avoidance

JEL classification: H2

1. Introduction

1.1. Why avoidance, evasion and administration are central, not peripheral, concepts in public finance

Most economic analysis of taxation presumes that tax liability can be ascertained and collected costlessly. As a description of reality, this is patently untrue. For example, in the U.S. the Internal Revenue Service (henceforth IRS) estimates that about 17% of income tax liability is not paid [1]; the figure for most other countries is probably higher. Furthermore, the resource cost of collecting what is paid can be large, in the U.S. probably about 10% of tax collections [2]. The tax structures themselves are undoubtedly skewed by the realities of tax evasion, avoidance, and administrative costs.

The standard models of taxation and their conclusions need to be modified in the light of these realities. Many practitioners of tax advice in developing countries believe that this change in emphasis is essential; for example, Casanegra de Jantscher (1990, p. 179) goes so far as to say that, in developing countries, "tax administration *is* tax policy" [3]. Bird (1983), Mansfield (1988), and Tanzi and Pellechio (1997) are useful summaries of the practical problems of the interaction of tax policy and tax administration in this context.

We believe that these issues are also critical in developed countries. In this setting, the issue is not the feasibility of certain taxes, but rather the optimality of alternative tax structures. For example, while in many developing countries an income tax that relies on self-reporting cannot be administered at all, in a developed country the question is to what extent optimal tax design should reflect the reality of evasion, the necessity of enforcement, and the costs of collection. In fact, tax systems do reflect these issues, although there is little systematic guidance offered by the academic public finance literature. The objective of this chapter is to collect and critique the now sizable literature that addresses these questions.

1.2. The evolution of tax structures

Scholars of the historical evolution of tax structure, notably Hinrichs (1966) and Musgrave (1969), have also stressed the importance of tax administration issues. They note that modern tax structure development has generally been characterized by a shift from excise, customs, and property taxes to corporate income and progressive

[1] Internal Revenue Service (1996).

[2] Slemrod (1996a).

[3] Others disagree. Groves (1974, p. 25) offers that: "Vetoing tax measures because of the difficulty of administering them is in most cases less compelling than doing so on the ground of their failure to conform to acceptable principles. Administration is usually amenable to improvement where violation of first principles is not. And administration of a given tax may often be improved most effectively in the process of attempting to administer it. The point is sometimes crucial in recommending taxes for so-called underdeveloped economies in our own time".

individual income taxes[4]. This shift has been made possible by the expansion of the market sector and relative decline of the rural sector, the concentration of employment in larger establishments, and the growing literacy of the population. Further changes in the technology of tax administration, including globalization and financial innovation, may now be pushing us away from progressive income taxes toward tax systems that rely more on broad-based consumption taxes such as the value-added tax (VAT), flatter rate structures for income taxation, or the "dual income tax" system recently adopted by certain Scandinavian countries, and described in Sørensen (1994).

Alt's (1983) treatment of the evolution of tax structure stresses the role of administrative and compliance costs. He argues that it has become increasingly easy to collect taxes from organized business rather than from households, and that one explanation for the widespread adoption of the VAT is that it imposes compliance costs without raising administrative costs, through incentives for self-policing. Kau and Rubin (1981) focus on changes in the cost of collecting taxes, and successfully relate growth of the U.S. federal government to reasonable correlates of collection cost, such as the literacy rate, the extent of female labor force participation, and the extent of the agricultural sector. Balke and Gardner (1991) contend that declining marginal collection costs can explain the stepwise growth in the size of government and the changes of taxation observed in the U.S. and U.K. They argue that major wars coincide with permanent improvements in tax instruments and tax collection technology, which facilitated permanent expansions in government size thereafter.

Putting aside the role of administrative issues in explaining the evolution of tax levels and tax structures, it is indisputable that these considerations are critical determinants of tax policy at a point in time. For example, an important set of generic aspects of income tax structure, such as the absence of taxation of imputed rents from consumer durables, taxation of capital gains (if at all) on a realization basis, and pre-set depreciation schedules, are undoubtedly largely driven by practical concerns of administerability. For these reasons, we believe that consideration of evasion, avoidance, and administration is essential to the positive and normative analysis of taxation. Our view corresponds closely to that of Blough (1952, p. 146):

> It is tax policy in action, not simply the wording of the statute, that determines how much the taxpayer must pay, and the effects of the payment. Knowledge of the statute is only a start in knowing a tax system. The interpretations placed on language by administrators and courts, the simplicity and understandability of tax forms, the competence and completeness of audit, the vigor and impartiality of enforcement, and the promptness and finality of action all influence the amount of revenue collected, the distribution of the tax load, and the economic effects of the tax.

[4] Although Hinrichs (1966) points out that tax structure development began with direct taxes rather than indirect taxes.

In this chapter we organize, explicate, and evaluate the modern literature that incorporates these considerations into the economics of taxation[5]. We do not claim to have put together a comprehensive survey of this literature, which is huge and multi-faceted, rather a guide to what we feel are the most important issues and contributions in this area.

1.3. Evasion, avoidance, and real substitution response

We begin with some definitions. The classic distinction between avoidance and evasion is due to Oliver Wendell Holmes, who wrote

> When the law draws a line, a case is on one side of it or the other, and if on the safe side is none the worse legally that a party has availed himself to the full of what the law permits. When an act is condemned as evasion, what is meant is that it is on the wrong side of the line ... *[Bullen v. Wisconsin (1916), 240. U.S. 625 at p. 630].*

Thus, the distinguishing characteristic of evasion is illegality[6]. In practice, of course, there are many gray areas where the dividing line is not clear, and sometimes the tax authorities may inappropriately characterize particular cases. One can draw a further distinction within the class of legal responses to taxation. At times we will refer to real substitution responses, or real responses for short, as those responses which come about because the tax law changes the relative price of different activities, and that induce taxpayers to respond by choosing a different consumption basket.

Conceptually distinct from real substitution responses are efforts to reduce one's tax liability without altering one's consumption basket, which we will refer to as avoidance. These are actions taken in response to the tax system that do not involve shifts along a given budget set. This definition covers a broad range of behaviors. One example is to pay a tax professional to alert one to the tax deductibility of activities already undertaken. Another example is to change the legal form of a given behavior, such as reorganizing a business from a C corporation to an S corporation[7], recharacterizing ordinary income as capital gain, or renaming a consumer loan as a home equity loan. A third example is tax arbitrage, when economically equivalent, but differentially-taxed, positions are held simultaneously long and short, producing tax savings. Finally, retiming a transaction to alter the tax year it falls under is an example of avoidance.

[5] Other surveys of these issues include Cowell (1990b), Andreoni, Erard and Feinstein (1998), Roth, Scholz and Witte (1989), and Alm (1999). Because of space constraints we have omitted any discussion of the role of tax practitioners, corruption in tax administration, tax amnesties, or business tax evasion.

[6] Kay (1980, p. 136) offers a different pair of definitions for evasion and avoidance: "Evasion is concerned with concealing or misrepresenting the nature of a transaction; when avoidance takes place the facts of the transaction are admitted but they have been arranged in such a way that the resulting tax treatment differs from that intended by the relevant legislation".

[7] Under U.S. tax law an S corporation retains the legal characteristics of a corporation, but is taxed as a pass-through entity such as a partnership. There are restrictions to becoming an S corporation, most notably on the maximum number of shareholders.

Fine distinction among the types of behavioral response to taxation is not possible and is for many issues not crucial. In general, changes in the tax structure will induce all the different kinds of response. Indeed, one of the goals of this chapter is to emphasize the common analytical aspects of issues that have traditionally been kept distinct.

1.4. General framework

Although there may be reasons, discussed later, for distinguishing among these categories of response to taxation, there is a common framework for analyzing these issues. Given the structure of the tax system and enforcement process, taxpayers are faced with opportunities to reduce their tax payments, or expected tax payments. There is a private cost to taking advantage of these opportunities, which may take the form of an altered consumption basket, an increasing probability of detection of, and penalty for, evasion, and/or a real resource cost of effecting avoidance or concealing evasion. This private cost depends on policies of the government that include, but are not limited to, the setting of tax rates and bases. The parameters of the tax administration and enforcement policies also matter, but these policies themselves are usually costly.

The tax system establishes the relative prices among this broad set of taxpayer activities. In the standard model, it establishes the relative price of leisure and other goods, as well as the relative price among the set of goods. In a more general framework it also sets the price of "honesty", meaning the incentives to evade, and establishes the cost and reward to legally reducing taxes via avoidance activities. The dimensions of taxpayer response interact. For example, real behavior may alter the cost of avoidance or evasion, thus changing the effective prices of real activities.

Although these are common themes, the literature to date has tended to isolate pieces of the overall problem. We follow that practice here, by beginning in Section 2 with a discussion of the now standard economic model of tax evasion[8]. Then, in Section 3, we introduce models that apply more generally to both evasion and avoidance. We then look at the empirical evidence, first in Section 4 about evasion, and then in Section 5 on avoidance. The remainder of the chapter addresses the implication for tax analysis of introducing these issues. Section 6 examines the fundamental issues of positive tax analysis, while Section 7 addresses normative issues. Section 8 concludes.

2. Theoretical models of evasion

2.1. The Allingham–Sandmo–Yitzhaki model

Suppose that the true tax base is known to the taxpayer, but is not costlessly observable by the tax collection agency. Then, under certain circumstances, the

[8] There is a vast literature which investigates non-economic perspectives on tax evasion. We do not have the space to discuss or evaluate this literature, and refer the interested reader to Roth, Scholz and Witte (1989).

taxpayer may be tempted to report a taxable income below the true value. In the seminal formulation of Allingham and Sandmo (1972) (henceforth A–S)[9], what might deter an individual from income tax evasion is a fixed probability (p) that any taxable income understatement will be detected and subjected to a proportional penalty (θ) over and above payment of the true tax liability itself. Later we introduce and discuss at length a "technology of evasion", in which evasion involves costs to the evader that might depend on the income and the amount of tax evaded.

In the A–S model, all real decisions, and therefore taxable income (y), are held fixed; only the taxpayer's report is chosen. The risk-averse taxpayer chooses a report (x), and thus an amount of unreported income y–x, in order to maximize expected utility:

$$\mathrm{EU} = (1-p)\,U(v+t(y-x))+pU(v-\theta(y-x)), \tag{1}$$

where v is true after-tax income, $y(1-t)$, t being the rate of (proportional) income tax. The von Neumann–Morgenstern utility function $U(\cdot)$ represents the individual's preferences toward risk. In this model the choice of whether and how much to evade is akin to a choice of whether and how much to gamble. Each dollar of taxable income understatement offers a payoff of t with probability $(1-p)$, along with a penalty of θ with probability p. If and only if the expected payoff to this gamble, $(1-p)\,t-p\theta$, is positive, every risk-averse taxpayer will chance some evasion, with the amount depending on the expected payoff and the taxpayer's risk preferences.

A critical issue, pointed out by Yitzhaki (1974), is whether the penalty for discovered evasion depends on the *income* understatement, as A–S assume, or on the *tax* understatement, as more accurately reflects practice in many countries. In the latter case, the maximand becomes $(1-p)\,U(v+t(y-x))+pU(v-\theta t(y-x))$, and the expected payoff per dollar of evaded income becomes $(1-p)\,t-p\theta t$. This is an important change, because it means that the tax rate has no effect on the terms of the tax evasion gamble; as t rises, the reward from a successful understatement of a dollar rises, but the cost of a detected understatement rises proportionately. The first-order condition for optimal evasion becomes

$$\frac{U'(y_{\mathrm{A}})}{U'(y_{\mathrm{U}})}=\frac{(1-p)}{p\theta}, \tag{2}$$

where y_{A} and y_{U} refer to net income in the audited and unaudited states of the world, respectively. Note that t does not appear in Equation (2), other than via an income effect in the definition of y_{A} and y_{U}. Compare this to the original A–S formulation, where t would be a multiplicative factor in the denominator of the right-hand side, implying that increases in t would proportionally increase the reward to getting away

[9] The Allingham–Sandmo paper applies to tax evasion the approach of the classic paper on the economics of crime by Becker (1968).

with understating income, but not proportionally increase the penalty, making evasion more attractive. Regardless of whether the penalty depends on the tax understatement or income understatement, more risk-averse individuals will, *ceteris paribus,* evade less. Individuals with higher income will evade more as long as absolute risk aversion is decreasing; whether higher-income individuals will evade more, as a fraction of income, depends on relative risk aversion. Evasion relative to income will decrease, increase or stay unchanged as a fraction of income depending on whether relative risk aversion is an increasing, decreasing, or constant function of income. Increases in either p or θ will decrease evasion.

Increasing t has both an income effect and, possibly, a substitution effect. If the taxpayer has decreasing absolute risk aversion, the income decline makes a less risky position optimal. An increase in t has a substitution effect, increasing the relative price of consumption in the audited state of the world, and thereby encouraging evasion, if the penalty is related to income, rather than tax avoided. In the latter case, if the penalty is related to the tax evaded, a tax increase has no substitution effect, so that an increase in t *reduces* evasion as long as there is decreasing relative risk aversion[10].

This simple version of the A–S model has been criticized on the grounds that it fails a simple reality check. If p is the fraction of returns audited in the U.S., about 0.015, and θ is the statutory penalty for non-criminal evasion, about 0.2, then based on the degree of risk aversion exhibited in other situations people should be evading a lot more than they apparently do. The intriguing question becomes why people *pay* taxes rather than why people *evade.* Much subsequent research, some of it surveyed below, has been addressed to reconciling the facts with the theory[11].

In the A–S model what limits the amount of evasion attempted is the taxpayer's risk aversion. At some point further evasion becomes just too big a gamble, so that at the chosen amount of evasion the marginal gain in expected tax savings is exactly offset by the marginal disutility of the extra risk taken on[12]. The model also predicts that a

[10] Note the similarity to the standard model of the effect of taxation on the optimal portfolio, in which a tax increase can increase the demand for the risky asset [Domar and Musgrave (1944)]. One difference is that, in a portfolio model, it is arguably inappropriate to ignore the effect of the tax scheme on the variability of government revenues [Gordon and Wilson (1989)]. This issue can be sidestepped in the context of a tax evasion model, because the "risks" are independent and therefore there is no social risk involved. It is important to distinguish the effect of a change in the environment on evaded income $(y-x)$ versus the impact on evaded tax liability, $t(y-x)$. With respect to changes in p and θ, there will be no interesting distinction. However, when t increases it is certainly possible that $(y-x)$ may decline at the same time $t(y-x)$ increases.

[11] One problem with this argument is that, for many types of evasion, the effective probablity of detection is much higher than the fraction of returns audited would suggest. For example, the p for non-reporting of wage and salary income subject to information reporting by employers is probably close to 1.0. Moreover, as long as several years of returns may be audited at once, the effective p may be several times higher than a one-year probabilty of audit would indicate.

[12] In the language of Kolm (1973), the evasion is accomplished by "the mere stroke of a pen". We consider below where the evasion is facilitated by supplying labor to an "underground" sector which offers better concealment possibilities.

risk-neutral individual would either remit no tax at all, or do no evasion, depending on whether evasion has a positive expected payoff. This "either–or" prediction is eliminated if the probability of detection is an increasing function of the amount of evasion, which is likely to characterize most tax systems. The implications of introducing an endogenous p depend on the precise relationship between p and evasion. For example, consider the case [discussed in Yitzhaki (1987)] where p is an increasing function of evaded income $(y-x)$. The risk-neutral taxpayer chooses x to maximize expected income,

$$\mathrm{EY} = ((1-p[y-x])(\eta+s)+p[y-x](\eta-\theta s)), \tag{3}$$

where $s \equiv t(y-x)$ is understated tax. If $p' \equiv \partial p/\partial(y-x)$ is positive, the first-order condition becomes

$$1-p-p\theta = p'(\theta+1)(s/t). \tag{4}$$

In this case, evasion will be constrained by the fact that p increases to offset what would otherwise be an increase in expected income.

The either–or prediction in the case of a risk-neutral taxpayer is also eliminated if there are distinct sources of income, each of which is subject to its own p. For example, employee labor income has a high p (due to information reporting by employers and computer matching), while "moonlighting" income has a much lower p. Faced with this situation, a risk-neutral individual reports all or none of each of the several sources of income, but may certainly report a fraction of total income [13].

The endogenous probability of detection can of course be applied to the case of a risk-averse taxpayer. In this case, at the margin the gain in expected value is offset by a combination of increased risk-bearing and an increased probability of detection. Cremer and Gahvari (1994) generalize this notion by introducing what they call a "concealment technology", which in our notation takes the form $p(y-x,((y-x)/y),m)$, where m represents taxpayer expenditure on concealment. The notion that the probability of detection can be increased by the taxpayer's expenditure is also present in Usher (1986), Kaplow (1990), Cowell (1990a), and Mayshar (1991).

2.2. *Jointness with labor supply*

Of particular interest is the relationship between the tax report decision and other consumer decisions. Most attention has been paid to labor supply, where the individual chooses how much labor to supply and how much labor income to report. The decision about how much income to report is made simultaneously with the decision of how much to work, so that it is impossible to adjust labor supply based on whether one

[13] As discussed in the next section, differential detection rates could also affect the sectoral supply of labor.

is caught evading. This problem may be posed as how much of a homogeneous labor income to report, which is equivalent to simultaneously choosing one's consumption basket and exposure to risk[14]. Models that belong to this group are based on extensions of the A–S model. Alternatively, the problem may be posed in the context of a model of the underground economy, in which there are two sectors with possibly different equilibrium wage rates and other different circumstances. The latter class of models allows for wage adjustment in response to policy changes, and thus are general equilibrium in nature.

In the extensions of the A–S model, the first-order condition for labor supply differs from that in a model without tax evasion only in that it contains mean marginal, instead of marginal, utilities. Whether mean marginal utility is bigger or smaller than the marginal utility depends on the sign of the second derivative of marginal utility, which is the sign of the third derivative of the utility function. On top of that, if utility is non-separable, the marginal utility functions depend on the sign of cross-derivatives, which further complicates the problem[15]. Baldry (1979) and Pencavel (1979) stress the difficulty of reaching any clear-cut comparative statics conclusions from such a model; the response of reported income to changes in tax rates, penalties, and fines becomes ambiguous. Thus, most models are based on particular restrictive assumptions about the utility function. For example, if the utility function is separable in consumption and leisure, then the marginal utility of leisure is independent of consumption. If, in addition, the marginal utility of consumption is linear (as in the function $U(C,L) = \alpha + \beta C + \gamma C^2 + \delta L$), the first-order condition for optimal labor supply is

$$(1-t)\,wU_1[wL+(1-p\theta)s] = U_2[L], \tag{5}$$

where s is the tax evaded and $(1-p\theta)s$ is the expected gain from evasion. Because evasion increases expected consumption for any given amount of leisure without changing the real wage, leisure would increase, and labor supply would decline. The real wage does not decline because the evasion opportunities are independent of the amount of work done. The critical importance of the relationship between the real consumption choices and the evasion or avoidance opportunities comes up again in the more general models discussed in Section 3. There we discuss cases where the avoidance opportunities do affect the real wage. In situations where labor income in the formal sector is reported by the employer to the tax enforcement agency as a matter of course, the only way to evade tax may be by "moonlighting" – working extra hours

[14] Models of this type resemble models of choice among risky occupations [e.g., Kanbur (1979)], except that in the latter the occupational choice is usually discrete, so that a "diversified" occupational portfolio is not allowed.

[15] Moreover, the conditions that the equilibrium investigated is on the increasing portion of the Laffer curve also depend on the curvature of the third derivative of the utility function, further complicating the issue.

at a different job – or by switching completely to the informal sector or "underground economy".

2.3. *Other uncertainty*

The basic model can also be extended to deal with other sources of uncertainty. Andreoni (1992) introduces a temporal nature to the tax evasion decision, recognizing the fact that the penalty for tax evasion, if detected, is assessed later than the tax saving. Andreoni deviates from the majority of the literature which assumes efficient market environments, and instead assumes that the taxpayer is constrained by credit rationing. Due to uncertainty, the income of the taxpayer fluctuates, as does the shadow price of income. Provided that non-monetary punishments are high enough to deter one from non-repayment of penalties and tax evaded, evasion may be viewed as a way of "borrowing" from the IRS. A constrained taxpayer may find it optimal to borrow when the shadow price of money is high enough during evasion and relatively low during repayment[16]. Andreoni models a situation where, in bad times, individuals evade as a way to smooth income streams; thus the IRS is a "loan shark". The conditional repayment of the loan occurs in a better state of the world.

Another aspect of uncertainty concerns the unpredictability of the tax liability itself, which arises when the "correct" tax liability is not clearly defined[17]. Uncertainty of true tax liability can be modeled by extending the Allingham and Sandmo framework. Scotchmer and Slemrod (1989) construct a model where, upon audit, the assessed tax liability is symmetrically centered around a known value with an equal probability of one-half. In this case the very concept of income understatement becomes problematic because the taxpayer is uncertain whether any given income declaration is correct or not.

There are now three possible outcomes that the taxpayer must consider. If the return is not audited (with probability $1-p$), true taxable income is irrelevant – the taxpayer merely pays the tax due on his declared taxable income. If the return is audited, there are two possible outcomes, depending on what the assessed tax liability turns out to be. Scotchmer and Slemrod (1989) show that increasing the dispersion of possible assessed taxable incomes induces increased compliance, given weak conditions about the taxpayer's attitudes toward risk. The intuition is that, for a given reported income, more dispersion lowers income in the least desirable state of the world, when the taxpayer is audited and his taxable income is determined to be the highest possible

[16] In this case the government may find it optimal to encourage tax evasion. The optimality of such a policy depends crucially on the non-existence of alternative methods of borrowing, including negotiated payment terms with the IRS, which can in some situations be arranged in the U.S.

[17] Long (1981) argues that the IRS exploits the unpredictability of tax liability to enhance its powers by using it as a license to decide cases in whatever way serves the government's interest at the time. She also notes that unpredictability makes the IRS's burden in providing criminal intent (rather than inadvertent errors) more difficult.

value. This increases the marginal utility of income in that state of the world, which is accomplished by increasing reported income and thus subjecting oneself to a lower penalty in the event this state of the world occurs. As long as the taxpayer exhibits declining absolute risk aversion, increasing the report is the optimal response.

Beck and Jung (1987) show that this conclusion may not hold when there is a continuous range of possible taxable income assessments. In this case one marginal benefit of increasing the income report is that it reduces the probability that a fine will be assessed. For a taxpayer reporting income below the mean of possible assessment, an increased dispersion of possible assessed incomes decreases the likelihood that the income report will be declared insufficient and a fine assessed, so that this component of marginal benefit is reduced. Thus, it is theoretically possible that increased dispersion will cause a lower report.

Note that uncertainty does not reduce tax evasion by as much as it reduces aggregate noncompliance in the sense of *true* aggregate tax liability minus tax paid. This is because one effect of uncertainty is to induce some taxpayers to pay *more* tax than they are legally obligated to pay, which reduces aggregate noncompliance but not the amount of individual tax evasion.

Scotchmer (1989) allows for the possibility that, by expending resources, the taxpayer can reduce the uncertainty of tax liability. The resources can be in the form of research by the taxpayer himself, or in the form of professional assistance hired. In this case the cost of unpredictability includes not only the disutility caused by uncertain tax liability but also the resources expended to reduce the uncertainty.

2.4. General equilibrium considerations

The A–S model and its direct descendants address only the demand for tax evasion by (potential) taxpayers. One might also consider the "supply" of evasion, and ponder the general equilibrium considerations of demand having to equal supply.

One context for this extension is the underground economy. Kesselman (1989) develops a set of models in which there are two sectors – above-ground and underground – which produce two distinct goods. Workers are homogeneous in their gross productivity in each sector of the economy (and in their consumption preferences), but must work only in one sector or the other. The workers, though, have differential distaste and risk aversion for tax evasion, and differential efficiency in concealment and other skills needed to operate successfully in the underground economy.

Although the precise results are model-dependent, three general conclusions obtain: (i) much of the gain from evasion may be shifted from the evaders to the consumers of output through lower prices, and the "marginal" evader gains nothing; (ii) relative price effects tend to dampen the impact of tax rate changes on the extent of evasion, and (iii) the effects of evasion on the marginal revenue response to tax rate changes will depend on consumers' elasticity of substitution between the sectoral outputs.

A key aspect of the foregoing model is that the act of tax evasion is tightly tied to the production of a distinct good. This need not be true, as is indicated by the simultaneous presence of above-ground and underground housepainters, repair people, and so on. Still, there is certainly evidence that evasion is concentrated in particular sectors, such as those that supply services directly to homeowners, because of the small scale of production that can aid concealment and the lesser need for receipts compared to services provided to businesses.

3. General models of avoidance and evasion

Because Allingham and Sandmo addressed tax evasion as a gamble, much of the subsequent literature focused on models in which taxpayers' risk aversion, and therefore higher-order characteristics of utility functions, play an important role. This focus has to some extent obscured other important aspects of the issue, such as the tax concealment technology, and also obscured the common aspects of evasion and what we have called avoidance. To highlight these issues we turn now to more general models of behavioral response to taxation.

Mayshar (1991) poses the taxpayer's problem as

$$\max_{X,S,L,Y} U(Y,L) \quad \text{subject to} \quad X = w[L - S - m(E)], \quad Y = X - T(X,S,E), \tag{6}$$

where X is output, S is sheltering effort, L is total labor effort, and Y is consumption. Mayshar labels $T(\,)$ the "tax technology"; it specifies the maximal taxes, T, collectible from a base X, when the tax authority selects a vector E of policy instruments, while the taxpayer devotes S in labor units to sheltering activity. It is reasonable to assume that $T_X > 0$ and $T_S < 0$ and, by construction $T_E > 0$. The function $m(E)$ represents unavoidable compliance costs associated with taxpaying, measured in labor units.

Although evasion as a gamble is not explicitly treated in this model, Mayshar argues that it can be presented in this framework; to do so S is defined as that certain payment which causes the same expected utility loss as the extra risk an evader takes on, for given expected tax payments. This forms the link between the A–S models of tax evasion and the models discussed in this section[18]. From the perspective of an A–S evasion model, $T_S < 0$ means that more evasion can lower expected tax payments, at a cost of more uncertainty.

Consider the first-order conditions with respect to L and S, respectively, where asterisks indicate an optimal value:

$$-U_L(Y^*,L^*)/U_Y(Y^*,L^*) = w[1 - T_X(X^*,S^*,E)], \tag{7}$$

$$w[1 - T_X(X^*,S^*,E)] \geqslant -T_S(X^*,S^*,E), \tag{8}$$

where Equation (8) holds as an equality if $S^* > 0$.

[18] Note that interpreting S, or C in the model of Slemrod (2001) discussed below, as the risk bearing cost of evasion, will impose restrictions on the $T(\,)$ or $C(\,)$ functions.

Expression (7) looks familiar: the marginal rate of substitution between consumption and leisure equals the net wage. But note that the effective marginal tax rate, $T_X(X^*, S^*, E)$, permits more complex marginal tax rates than the standard linear tax model, where $T(X^*, S^*, E)$ would equal tX^*, and so T_X would equal t. In Expression (7), the effective marginal tax rate may depend on the sheltering activity of the taxpayer and/or the policy instruments of the government, interpreted more broadly than simply announcing a tax schedule. Expression (8) states that, because sheltering is accomplished by using labor, at an interior optimum its opportunity cost $w(1 - T_X(\ ))$ will be equal to its marginal private gain, which is the marginal tax saving, $-T_S$.

Slemrod (2001) investigates a related model in which the private cost of achieving reductions in taxable income (denoted A, for income avoidance) is $C(wL, A)$, where wL is true labor income; he argues that, in general, $C_1 < 0$ and $C_2 > 0$[19]. If we imbed this avoidance technology into the taxpayer choice under a linear income tax, the maximization problem becomes

$$\max_{L,A} U(Y, L), \quad \text{subject to} \quad Y = w(1-L) - t(w(1-L) - A) - C(wL, A). \tag{9}$$

Before pondering the general implications of this formulation, first consider the special case where $C(wL, A) = C(A)$. In this case the first-order condition for labor supply is identical to the standard model without avoidance. The first-order condition for A is simple and straightforward, $C' = t$, implying that avoidance ought to be pursued until its marginal cost equals its marginal saving in tax liability, equal to t. In this situation a tax rate hike unambiguously increases A. Furthermore, its effect on L is no different than in the standard model, except to the extent that the income effect is altered by the possibility of avoidance.

The story is enriched when the avoidance, or tax, technology becomes $C(wL, A)$. The effective marginal return to working becomes $w(1 - t - C_1)$, where $-wC_1$ is a subsidy to working that Slemrod (2001) dubs the "avoidance-facilitating" effect; for example, a given level of allegedly work-related deductions looks more plausible if it is taken against a larger gross income. The term $(t - C_1)$ is analogous to T_X in Mayshar's model, and makes explicit how the avoidance technology influences the incentive to supply labor.

Several insights emerge from this modeling of the tax environment. First of all, the substitution effect of labor supply does not respond identically to the two components of the statutory after-tax wage rate, w and $(1 - t)$. Changes in $(1 - t)$ trigger avoidance responses which are not triggered by changes in w. While both labor supply and avoidance respond to both w and $(1 - t)$, they do not do so symmetrically. This

[19] Slemrod (2001) is less general than Mayshar in that it presumes a flat-rate statutory tax system; it does not presume that tax sheltering or avoidance must be "produced" with the taxpayer's own time. One superficial difference is the adoption by Slemrod of a cost function approach to avoidance, compared to Mayshar's production function for tax receipts.

implies that econometric studies of labor supply (and avoidance) ought to differentiate responses to w and $(1-t)$. Furthermore, one should not conclude, as does Rosen (1976), that a differential response to w and $(1-t)$ necessarily represents "taxpayer illusion"[20]; instead it could be reflecting the avoidance technology.

Mayshar and Slemrod addressed the possibility that changes in the tax system will induce from taxpayers all three types of behavioral response. For example, an increase in the rate of a proportional income tax will provide an incentive to substitute leisure for goods, to (depending on the penalty structure) increase evasion, and increase avoidance. Other interactions among the three types of behavioral response have been investigated, as well. Cowell (1990a) develops a model in which the taxpayer can evade, but can also legally shelter income for a fixed cost Γ and a constant marginal cost γ, where $\gamma < t$. These cost assumptions generate the result that if an honest (or highly risk-averse) person shelters any of his income (Y), then he will automatically shelter all of it, and will do the latter if $\Gamma + \gamma Y < tY$. Cowell then investigates whether sheltering will co-exist with evasion, and asserts that the optimum is *not* characterized by an equality between the marginal cost of avoidance and evasion. This is because sheltering reveals to the tax authority that the taxpayer's true income must be at least $\Gamma/(t-\gamma)$. He argues that there may be a class of shelterers who would also have been evaders, had it not been for the attention drawn by sheltering, and that in some cases there may be a complete polarization between evaders and avoiders.

In Cross and Shaw (1982), taxpayers must make expenditures to learn about and (in the case of avoidance) document both avoidance and evasion activities[21]. Two avenues of interaction arise. First, in a progressive tax system, expenditure on avoidance or evasion reduces the marginal tax rate, thus reducing the return to engaging in the other[22]. Second, investment in avoidance may reduce the marginal cost of evasion, or vice versa. For example, while investigating an illegal but undetectable "tax shelter", a (barely) legal tax shelter arrangement may be uncovered without much additional investment of time.

4. Descriptive analysis of evasion and enforcement

4.1. The extent of tax evasion

4.1.1. Data problems

Ascertaining the extent and characteristics of evasion immediately runs into two problems – one conceptual and one empirical. The conceptual problem is that, although

[20] Although, note that in his empirical analysis Rosen (1976) does not detect a significant differential response.

[21] In some situations, more evasion may be associated with less cost. For example, not bothering to trace a miscellaneous source of income is less costly than tracking down whatever receipt or Form 1099 would document the income. Not filing a return at all happens to minimize compliance cost.

[22] Alm (1988) also examines the simultaneous choices of evasion, avoidance and reported income, and investigates the effects of other fiscal variables on these choices.

one can assert that legality is the dividing line between evasion and avoidance, in practice the line is often blurry. Sometimes the law itself is unclear, sometimes it is clear but not known to the taxpayer, sometimes the law is clear but the administration effectively ignores a particular transaction or activity. The importance of these factors certainly differs across situations.

The other difficulty is that, by its nature, tax evasion is not easy to measure – merely asking just won't do. Several different approaches have been attempted. One approach relies on inferring the level or trends in noncompliance from data on measurable quantities, such as currency demand or national income and product accounts. The monetary indirect estimates are based on the presumption that most unreported economic activity takes place in cash, and that some time in the past the underground economy was small. In Gutmann (1977), increases in the ratio of currency to demand deposits since 1937–41 measure the underground economy; in Feige (1979), changes since 1939 in the ratio of total dollar transactions to official GNP since 1939 measure it. Tanzi (1980) estimates regressions explaining the ratio of currency to money defined as M2, and interprets the portion explained by changes in the tax level as an indication of changes in the size of the underground economy. None of these approaches is likely to be reliable, however, as their accuracy depends either on unverifiable assumptions or on how well the demand for currency is estimated. The indirect noncompliance estimates based on discrepancies between national accounts measures of income and income reported to the tax authority are also problematic. For one thing, national income estimates of several key forms of income are based on tax return data. Second, there are many inconsistencies between how income is defined for tax purposes and for national accounts. However, Engel and Hines (1999), in a study of tax evasion dynamics which focuses on the possibility of retrospective examination of previous-years' returns, study this measure of evasion in the U.S. for the years 1947 to 1993 and find that it responds as their model predicts. For example, annual fines and penalties imposed by the IRS subsequent to audits are correlated with contemporaneous and several lags of tax evasion as calculated from national income statistics.

The most reliable source of information about tax compliance concerns the U.S. federal income tax, and exists because of the IRS's Taxpayer Compliance Measurement Program, or TCMP. Under this program, approximately every three years from 1965 until 1988 the IRS conducted a program of intensive audits on a large stratified random sample of tax returns, using the results to develop a formula used to inform the selection of returns for regular audits. The TCMP data consist of line-by-line information about what the taxpayer reported, and what the examiner concluded was correct. This data formed the basis for the IRS estimates of the aggregate "tax gap", and provides much useful information about the patterns of noncompliance with respect to such variables as income, occupation, line item, region of the country, age, and marital status. While informative, it is widely recognized that even the intensive TCMP audits imperfectly reveal particular kinds of noncompliance, such as income from the underground economy.

4.1.2. Patterns of noncompliance

We cannot adequately review here what is known about the extent and nature of tax evasion for all taxes in all countries at all times. Rather, in what follows we offer a few salient facts about the recent U.S. income tax, mostly gleaned from the TCMP data just discussed.

(1) With audit coverage hovering at about 1% and an extensive information reporting and matching program, evasion is estimated to be 17% of true tax liability[23].
(2) The extent of evasion varies widely across types of gross income and deductions; for example, the 1988 TCMP reports that the voluntary reporting percentage was 99.5% for wages and salaries, but only 41.4% for self-employment income (Schedule C). These percentages clearly correlate positively with the likelihood of income understatement being detected.
(3) Evasion (as measured by underreported income, not tax liability), rises with income but at a less than proportionate rate. Christian (1994) reports that in 1988, taxpayers with (auditor-adjusted) incomes over $100 000 on average reported 96.6 percent of their true incomes to the IRS, compared to just 85.9 percent for those with incomes under $25 000[24].
(4) Within any group defined by income, age, or other demographic category, there are some who evade, some who do not, and even some who overstate tax liability[25]. For example, of middle-income (auditor-adjusted income between $50 000 and $100 000) taxpayers in 1988, 60% understated tax, 26% reported correctly, and 14% overstated tax [Christian (1994, p. 39)].

4.2. Determinants of evasion

Empirical attempts to more systematically establish how compliance responds to aspects of the tax environment have met with limited success, primarily due to the data problems discussed in Section 4.1.1[26]. Three approaches dominate the literature[27].

[23] It is probably higher in most other countries. For example, Alm, Bahl and Murray (1991) put the figure (for avoidance and evasion) at 46% for the Jamaican income tax of 1983. Richupan (1984) cites studies of tax evasion in developing countries indicating that it is not uncommon for half or more of potential income tax to be uncollected.

[24] One explanation for this pattern is almost certainly that tax returns of high-income households are more likely to attract IRS attention. Another potentially important factor is that the TCMP results do not account for the noncompliance of business entities, which are more germane for higher-income individuals.

[25] Note that Erard (1997) concludes that a large fraction of noncompliant reports may be unintentional.

[26] A former colleague, Harvey Galper, once put the problem this way: "Regression analysis of tax evasion is straightforward, except for two problems: you can't measure the left-hand side variable, and you can't measure the right-hand side variables!"

[27] In addition to the econometric methodologies discussed below, laboratory experiments typically involving students engaged in a multi-period reporting game, have been employed. [See, for example,

4.2.1. Cross-sectional analysis

Clotfelter (1983) was the first attempt to make use of the TCMP data to investigate how noncompliance responded to changes in the environment. He estimated a tobit model, explaining, for each of ten audit classes, noncompliance as a function of the combined federal and state marginal tax rate, after-tax auditor-adjusted income, and a set of demographic variables available on tax returns. The most striking conclusion is that noncompliance is strongly positively related to the marginal tax rate, with the elasticity ranging from 0.5 to over 3.0. This finding is apparently consistent with the basic A–S model, but not with the extension proposed by Yitzhaki.

Beron, Tauchen and Witte (1992) investigate TCMP data aggregated by the IRS to the three-digit zip code level. They find that increasing the odds of an audit significantly increases reported AGI and tax liability for some, but not all, of the groups. In an attempt to deal with the potential endogeneity of the intensity of enforcement, they model the simultaneous determination of tax reporting and the log odds of an audit for each of the several audit classes in each zip code area. Their instrument for this is the level of IRS resources relative to the number of returns[28]. Although Beron, Tauchen and Witte argue that it is a valid instrument because the IRS has not been able to distribute its resources among districts so as to achieve its goals, this is not convincing: it is reasonable that the IRS attempts to target its resources toward areas believed to be particularly noncompliant, thus invalidating use of IRS resources as an instrument.

Subsequent studies have produced mixed results. Of particular interest is work by Feinstein (1991), who performed a pooled cross-section analysis of 1982 and 1985 TCMP data, thus mitigating the problem that in a single cross-section (other than for cross-state differences) the marginal tax rate is a (complicated, non-linear) function of income, making it difficult to separately identify the tax and income effect. Feinstein's analysis suggests a negative impact of the marginal tax rate on evasion, which contradicts Clotfelter's results but is consistent with the A–S model as adjusted by Yitzhaki.

Klepper and Nagin (1989) investigate the characteristics of evasion across line items, and find that noncompliance rates are related to proxies for the traceability, deniability,

Baldry (1987) and Alm, Jackson and McKee (1992)]. These results are subject to the canonical criticisms of laboratory studies: that the setting is artificial, and the participants are not demographically similar to those making the actual decisions, and therefore do not come to the decision problems with the same array of experiences and expectations about the environment. These criticisms may be especially salient in this context, because the experiments differ from general problems in choice under uncertainty only by the labeling of the choice as having to do with taxes, and as compliant or not rather than gambling or not.

[28] Dubin and Wilde (1988) perform a similar analysis on the zip-code aggregated data, and use the same instrument. They defend this choice by claiming that, in an analysis of the time path of state-level IRS budgets, they were found to be independent of compliance levels, and predominantly determined by the share of total returns filed.

and ambiguity of the items, which are in turn related to the probability that evasion will be detected and punished. They also find evidence of a "substitution effect" across line items, such that greater noncompliance on one item lowers the attractiveness of noncompliance on others, because the latter jeopardizes the expected return to the former by increasing the probability of detection.

4.2.2. Time-series analysis

Dubin, Graetz and Wilde (1990) make use of state-level time series cross-section data from 1977 through 1986 to investigate the impact of audit rates and tax rates on tax compliance. They do not, though, have a direct measure of noncompliance, but instead use tax collections per return filed and returns filed per capita as (inverse) measures of noncompliance. They conclude that the continual decline in the audit rate over this period caused a significant decline in IRS collections – amounting to $41 billion by 1985.

4.2.3. Controlled experiments

As discussed above, analysis of both cross-section and time-series historical data is subject to severe difficulties of measuring the parameters of the environment and in knowing the source of any variation in these parameters. Controlled experiments can avoid all of these problems, but, for cost and other implementation reasons, are rare.

One recent exception is reported by Slemrod, Blumenthal and Christian (2001), in which the State of Minnesota Department of Revenue conducted a randomized controlled experiment with respect to four aspects of the tax compliance environment: the threat of an audit, the provision of special return preparation information services, moral appeals, and a redesigned tax form. With regard to the first, they find that, for low- and middle-income taxpayers, a threat of certain audit[29] produced a small, but statistically significant, increase in reported income, which was larger for those with greater opportunities to evade[30]. However, for high-income taxpayers the audit threat was associated with on average a *lower* income report. The authors speculate that sophisticated, high-income, taxpayers view an audit as a negotiation, and view reported taxable income as the opening (low) bid in a negotiation which does not necessarily result in the determination and penalization of all noncompliance. Based on the same experiment, Blumenthal, Christian and Slemrod (2001) find no evidence that either of two written appeals to taxpayers' consciences had a significant effect on aggregate compliance.

[29] The audit threat was delivered by letter in January following the tax year.

[30] The approach is a "difference-in-difference" analysis; that is, the increase in reported income over the previous year of the treatment group is compared to the increase in reported income of the control group.

5. Descriptive analysis of avoidance

5.1. Dimensions of avoidance

Stiglitz (1985) distinguishes three basic principles of tax avoidance within an income tax: postponement of taxes, tax arbitrage across individuals facing different tax brackets (or the same individuals facing different marginal tax rates at different times), and tax arbitrage across income streams facing different tax treatment. Many tax avoidance devices involve a combination of these three principles. In an example used by Stiglitz, the basic feature of an Individual Retirement Account (IRA) is the postponement of tax liability until retirement; if the individual faces a lower tax rate at retirement than at the time the income is earned, then the IRA also features tax arbitrage between different rates. Finally, if the individual can borrow to deposit funds in an IRA and the interest incurred to finance the deposit is tax deductible, then the IRA is a tax arbitrage between two forms of capital, one of which is taxed, and the other of which is not taxed[31]. Stiglitz argues that, with perfect capital markets, these three principles can be exploited to eliminate all taxes while leaving the individual's consumption and bequests unchanged relative to the zero tax case, and facing no more risk than in the original situation. But capital markets are not perfect, and therefore all tax liability is not eliminated by tax avoidance[32], and to reduce tax liabilities distorting actions (such as investment in sectors where it is easier to convert ordinary income into capital gains) are utilized. There is considerable empirical evidence testifying to the extent and tax sensitivity of these kinds of avoidance behavior.

5.1.1. Retiming

There is abundant support for the notion that the timing of certain transactions can be extraordinarily responsive to changes in tax rates. Perhaps the most striking example was the response of capital gains realizations to the tax rate increase scheduled to occur on January 1, 1987, but fully anticipated by the fall of 1986. Aggregate realizations in 1986 were *twice* what they were in any previous year or for several years thereafter. As Burman, Clausing and O'Hare (1994) document, capital gain realizations on corporate stock in December of 1986 were seven times higher than in the previous December. Another striking example of timing response is provided by Goolsbee (2000), who documents that, in advance of the expected 1993 increase in the U.S. top individual

[31] The IRA example makes clear that in certain cases (some of) the avoidance behavior is the result of a conscious tax policy choice, in this case with the intent of increasing saving. Another excellent example is capital gains, where taxation upon realization rather than accrual allows for deferral of tax liability, often into periods of lower taxation, and where gains are completely excused from taxation at death due to the step-up of tax basis.

[32] There are also policy responses to avoidance, such as limits on loss offsets and interest deductions.

tax rate, corporate executives realized a huge amount of income in 1992, primarily through exercising non-qualified[33] stock options.

Sophisticated econometric techniques using panel data have been developed for separately identifying the timing responses to tax rate changes over time from the permanent behavioral response to a changed tax rate. These new techniques have been applied to both capital gains realizations [Burman and Randolph (1994)] and charitable contributions [Randolph (1995)]. In both cases the results suggest that the retiming effect dominates the permanent effect.

5.1.2. Tax arbitrage

Tax arbitrage activity takes advantage of inconsistencies in the tax law, featuring economically offsetting positions which have asymmetric tax treatments. Examples range from sophisticated derivative financial instruments to the more mundane cases of doing tax-deductible borrowing to finance tax-deferred IRA contributions or tax-exempt bond purchases.

5.1.3. The classification of income

The classic example of income reclassification, also termed income shifting, is turning ordinary capital or labor income into preferentially-taxed capital gains. In another example, Maki (1996) and Scholz (1994) have documented that, following the Tax Reform Act of 1986, there was a large shift from no-longer-deductible consumer interest into still-deductible mortgage or home equity loans. There is anecdotal evidence that, following the introduction of the R&D credit in the United States, much business activity was "discovered" to have a significant research component. Gordon and MacKie-Mason (1990, 1997) have investigated how, when the Tax Reform Act of 1986 lowered the top individual rate below that of the corporate rate, there was a large shift from C corporations into S corporations, which are taxed like partnerships and therefore are not subject to the corporation income tax. Gordon and Slemrod (2000) discuss the shifting of income between the corporate and individual tax base via the method of compensation, and document evidence of such shifting in the United States.

5.2. The extent of avoidance

No one has attempted to calculate for avoidance a counterpart to the aggregate evasion "tax gap". There is, though, some indirect evidence that the avoidance tax gap is

[33] For non-qualified stock options, the difference between the exercise price and the issue price is taxable at ordinary income tax rates at the time of exercise, and is deductible from the employer's taxable income at the same time.

large. Gordon and Slemrod (1988) calculated that the U.S. tax system of 1983 raised approximately zero revenue from taxing capital income, due to the combination of legislated deviations from a pure income tax and tax arbitrage[34]. As to the incidence of the avoidance opportunities, Agell and Persson (1990) and Gordon and Slemrod (1988) argue that the availability of tax arbitrage opportunities will generally benefit those at the bottom and top of the tax rate distribution, to the disadvantage of those in the middle. This generally corresponds to low- and high-income individuals, respectively, but there are exceptions to that rule; high-income individuals benefit through their ownership of tax-preferred pension assets.

6. Fundamentals of tax analysis

Having completed a review of the positive, or descriptive, analysis of tax evasion and avoidance, we turn now to the normative analysis of taxation. However, before we proceed to that task, we must first reconsider the fundamental building blocks of tax analysis – the evaluative criteria of equity and efficiency – to check whether these concepts need to be revised.

6.1. Equity

6.1.1. Vertical equity

Analyses of the distributional impact of taxation, especially those based on tax return data, ought to account for the presence of evasion. The evidence cited in Section 4.1 – that noncompliance as a fraction of true income *declines* with true income – suggests that standard analyses of incidence based on the statutory rates and base may understate the progressivity of the tax burden[35]; Bishop, Chow, Formby and Ho (1994) find this for the United States using the 1985 TCMP data, although Alm, Bahl and Murray (1991) reach the opposite conclusion about Jamaica[36].

6.1.2. Horizontal equity

Horizontal equity – the idea that equals should be treated equally by the tax system, or that tax liability should not depend on any of a set of irrelevant characteristics –

[34] It is likely that the Tax Reform Act of 1986 mitigated the avoidance tax gap by reducing the dispersion of marginal tax rates and tightening up the rules about tax arbitrage behavior.

[35] This evidence also suggests that tax return data may overstate the inequality in the distribution of incomes. Because the data on tax evasion are flawed, one should keep in mind that the theoretical arguments discussed in Section 5.2 imply that data on reported incomes will *understate* the true dispersion of income.

[36] A complete incidence analysis would account for the costs borne by evaders in the form of exposure to risk and concealment expenses, neither of which is accounted for in the studies mentioned.

is central to an assessment of the impact of tax avoidance and evasion. To see this, compare two tax situations, one in which there is a linear income tax rate of 20% and everyone reports their true income, and another in which the tax rate is 40% and everyone (costlessly) reports exactly half their income. In this case the two systems are identical with respect to both horizontal and vertical equity. Now imagine that, in the second system, on average everyone reports half their income, but that the fraction differs systematically by income. In that case replicating the progressivity of the first tax system will require a more complicated, non-linear, system of rates. If, however, evasion varies *within* income classes, no revision of the tax rate schedule can compensate, and there will be horizontal inequity.

In the context of the A–S model of tax evasion, the horizontally inequitable tax burden will depend on the taxpayer's degree of risk aversion. Less risk-averse households will gain more from the availability of a gamble with given positive expected value. In contrast, common parlance would ascribe any horizontal inequity to variations in honesty, with the honest, or dutiful, citizens left holding the bag by the dishonest. In the typical economic model, though, there are no honest or dishonest individuals, only utility-maximizers; thus, this distinction can be introduced only artificially by simply positing that some individuals do not pursue tax evasion. The same kind of artificial differentiation across people can be made with regard to tax avoidance by positing that some people have an aversion to such behavior; as Steuerle (1985, p. 78) says: "Some taxpayers simply do not enjoy playing games no matter what the certainty of the return; the U.S. tax system is designed to insure that such individuals pay a greater share of the tax burden than those who are not so hesitant". Steuerle (p. 19) concludes that "taxpayers pay unnecessary taxes because of the simplicity of their filing response or their lack of knowledge of the tax laws".

6.1.3. Incidence

The theory of tax incidence – who bears the burden of a given tax structure – begins with three basic principles: (i) the burden of all taxes must be traced back to individuals; (ii) individuals with relatively elastic demand (or supply) of a taxed good tend to escape the burden of tax imposed on that good; and (iii) in the long run the incidence of a tax levy does not depend on which side of the market bears the legal responsibility for remitting the tax to the government. Introducing avoidance and evasion preserves the methodological importance of the first two principles[37], but calls the third into question. A complete analysis of the incidence of a particular tax requires specifying the remittance process and positing an avoidance technology for both the suppliers and demanders of the taxed good.

Avoidance opportunities alter the analysis of incidence for two separate reasons. First, their presence affects the behavioral response to a change in the tax system,

[37] Section 2.4 discusses some models of tracing the incidence of tax evasion.

and this alters what otherwise would be the change in equilibrium prices. Second, the presence of avoidance alters the link between tax-inclusive prices and welfare. This suggests that the incidence (not to mention the efficiency) of a tax may depend on which side of the market the responsibility for remittance falls. That is in stark contrast to the standard model, under which that is irrelevant to the long-run incidence[38].

6.1.4. Are changes in the social welfare function necessary?

In models with heterogeneous citizens, the standard objective function is a social welfare function which has as arguments the utility level of each citizen – accepting the individuals' own relative valuations of goods and services – where the shape of the social welfare function implicitly determines the social value placed on the distribution of utilities as opposed to the sum of utilities. In the presence of uncertainty, the *expected* utilities of individuals are the relevant arguments – accepting the risk preferences of consumers. Cowell (1990b) questions the appropriateness of according the same social weight to investigated and guilty taxpayers as is applied to the innocent or uninvestigated, and argues that there may be a case for putting a specific discount on the utility of those "who are known to be antisocial" (p. 136). Cowell investigates a few alternative social objective functions, including one in which any private benefit derived from the proceeds of evasion is assigned a social weight of zero, but in our opinion no convincing alternative that provides reasonable policy prescriptions has yet been presented.

6.2. A taxonomy of efficiency costs

In the standard model the efficiency cost of taxation is entirely due to the fact that, because of the change in relative prices, individuals are induced to select socially suboptimal consumption baskets – to substitute away from relatively highly-taxed goods to relatively lightly-taxed goods, such as leisure. A standard exercise in optimal taxation theory is to describe the tax system that minimizes these costs, or to describe the tradeoff between these costs and the distribution of welfare in the society.

In the presence of avoidance and evasion, a broader concept of efficiency cost is needed. In what follows, we describe and comment on three additional components of the social cost of taxation and discuss the problems that arise in introducing these costs into formal models of optimal taxation.

6.2.1. Administrative costs

Tax administrations deal, among other things, with information gathering. But this is a difficult element to model because information varies in quality. There is a qualitative

[38] There are exceptions. Consider, for example, the debate between Tanzi (1992) and Dixit (1991) over the implications of tax collection lags for the optimal amount of inflationary finance. Tanzi (1992) argues that, when consumption taxes are collected by firms in advance and held by them for the duration of the collection lag, inflationary finance implies a real redistribution of income from consumers to sellers.

difference between an auditor "knowing" that a given taxpayer is evading and having sufficient evidence to sustain a court finding to that extent. Also, the cost of gathering information depends on how accessible the information is, and whether it can be easily hidden. There are several advantages to taxing a market transaction relative to taxing an activity of the individual such as self-consumption. First, in any market transaction there are two parties with conflicting interests. Hence, any transaction has the potential of being reported to the authorities by one unsatisfied party. A second property is that the more documented the transaction, the lower is the cost of gathering information on it. For this reason it is easier to tax a transaction that involves a large company, which needs the documentation for its own purposes, than to tax a small business, which may not require the same level of documentation. Finally, market transactions establish arms-length prices, which greatly facilitate valuing the transaction. Administrative cost may also be a function of the physical size and the mobility of the tax base (it is harder to tax diamonds than windows), whether there is a registration of the tax base (e.g., owners of cars, holders of drivers' licenses), the number of taxpayer units, and information sharing with other agencies[39]. It is also an increasing function of the complexity and lack of clarity of the tax law.

Administrative costs possess two additional properties that complicate the modeling of tax administration issues: they tend to be discontinuous and to have decreasing average costs with respect to the tax rate. To see the first property, consider two commodity tax rates, denoted by t_1 and t_2. If $t_1 = t_2$, then only the total sales of the two commodities need be reported and monitored. If, however, the two rates differ even slightly, then the sales of the two commodities must be reported separately, doubling the required flow of information. There are decreasing average costs because the cost of inspecting a tax base does not depend on the tax rate (except to the extent that people are more inclined to cheat with a higher tax rate). Hence, a higher tax rate reduces the administrative cost per dollar of revenue collected [Sandford (1973)]. Administrative cost may also be a function of the combination of the taxes employed and their rates, because the collection of information concerning one tax may facilitate the collection of another tax (e.g., inspection of VAT receipts may aid the collection of income tax).

6.2.2. *Compliance costs*

Slemrod (1996a) estimates that, for the U.S. income tax, the private compliance cost is about 10 cents per dollar collected. Sandford (1995) presents estimates for a variety of taxes in several countries. Some of that cost is an unavoidable cost of complying with the law, and some of it is voluntarily undertaken in an effort to reduce one's tax bill, but in either case it approximately represents resource costs to society. In almost all cases the *private* compliance costs dwarf the *public* administrative costs of collecting taxes,

[39] A good description of the properties of administrative cost can be found in Shoup, Blough and Newcomer (1937, pp. 337–551).

which the IRS estimates at 0.6 cents per dollar collected for all the taxes it administers. Integrating compliance costs into formal models in a meaningful way is tricky. As an example of the modeling difficulties this topic poses, consider the following problem: when is it optimal to delegate to employers the authority to collect taxes and convey information about employees, thus requiring the administration to audit both the taxpayer agent and the taxpayer himself, and when is it optimal to deal only with the employee? Clearly, given that the employer already has the necessary information, it would save administrative costs to require him to pass it along to the tax administrator. This might also reduce total social costs if the cost of gathering information by the administration is higher than the increase in cost caused by imposing a two-stage information-gathering system[40].

However, the potential efficiency of involving taxpayers in the administrative process must be tempered with a practical consideration. Administrative costs must pass through a budgeting process, while compliance costs are hidden. Hence, there may be a tendency to view a policy which reduces administrative cost at the expense of an equal (or greater) increase in compliance costs as a decrease in social cost, because it results in a decrease in government expenditures. We will discuss this issue further in Section 7.

6.2.3. The risk-bearing costs of tax evasion

In the Allingham–Sandmo model, tax evasion occurs only if the taxpayer expects to increase his expected income by evading taxes, including the expected fines that he would have to pay if he were caught; it continues until, at the margin, the increased expected income is offset by the increased risk-bearing. Hence, a taxpayer who evades taxes increases both his exposure to risk and his expected income. This additional exposure to risk is a deadweight loss to society. In principle, the taxpayer could be better off under an agreement whereby the taxpayer pays at least as much as the government currently collects, while the government ceases to audit. Assuming a risk-neutral government, the risk-bearing cost of tax evasion is equal to the risk premium that the taxpayer would be ready to pay in order to eliminate the exposure to risk [Yitzhaki (1987)]. Depending on the other assumptions about the probability of detection, the penalty structure, and risk aversion, the risk-bearing costs of evasion may be a continuous function that increases with the tax rates. These costs are in addition to the compliance costs voluntarily incurred by an individual attempting to minimize the expected cost by camouflaging the evasion or shifting to an otherwise less remunerative occupation.

[40] Note that a withholding system requires two information gathering systems and might generate incentives for the withholding agent to evade the taxes it collects, or to collaborate with withholdees in withholding less than required [Yaniv (1988, 1992)]. In a period of rapid inflation, the gain to the agent from withholding may exceed the cost.

7. Normative analysis

7.1. Optimal tax administration and enforcement

Avoidance and evasion pose two challenges for the normative analysis of taxation. First, they introduce a new set of policy instruments whose optimal setting is at issue. These include the extent of audit coverage, the penalty imposed on detected evasion, and the structural integrity of the tax code itself, which determines the extent and nature of avoidance opportunities. Second, they invite a rethinking of standard taxation problems, including the optimal setting of commodity tax rates and optimal progressivity.

7.1.1. Optimal penalties

Consider the A–S model of a representative consumer whose true income is exogenous and whose only choice concerns how much of that income to report. This choice depends on two policy instruments set by the government, p, which has a resource cost due to the need for auditors and the related infrastructure, and θ, which is a fine for detected evasion, which is a transfer with no resource cost.

It has been well known since Becker (1968) that in this setting a government concerned with maximizing the *ex ante* utility of its representative citizen will want to set θ as high as possible, allowing p to be as low as possible. This policy of "hanging violators with a probability of zero" deters evasion while minimizing the resource cost of the deterrent – p represents a real resource cost but θ is simply a transfer. But this kind of model ignores, *inter alia,* the possibility of a corrupt tax administrator who abuses the system or, alternatively, harshly punishes someone who commits an honest mistake[41]. The harsher the penalty, the more damage that can be inflicted by a corrupt administrator or, in the case of an honest mistake, the more capricious the system is. Hence, the harsher the penalty, the more detailed and cautious the prosecution process should be, although this may increase its administrative costs. In the absence of modeling the interaction between the penalty rate and administrative costs, analytical models usually assume a ceiling on the penalty rate.

7.1.2. Optimal randomness

Auditing some taxpayers and not others inevitably introduces some *ex ante* uncertainty and some *ex post* horizontal inequity. This suggests a link to an earlier literature in public finance, in which Stiglitz (1982) and Weiss (1976) each argued that, even in a world of risk-averse citizens, it may be optimal for the government to introduce some randomness into its net tax (or transfer) to individuals. The argument depended on

[41] Polinsky and Shavell (2000) examine this and other issues involved in the optimal setting of penalties for crime including but not restricted to tax evasion.

the second-best nature of the problem, in which an income tax distorted the labor–leisure choice. For some utility functions, Stiglitz and Weiss argued, the introduction of random payments induced people to work harder, thus mitigating the labor market distortion; in some cases the value of the increased labor more than offset the utility loss from the randomness introduced.

This argument has clear implications for the optimal enforcement of the income tax, because it suggests that one of the presumed social benefits of greater enforcement – the reduced uncertainty of payment of a given expected value of taxes – may be mitigated by the increased labor supply distortion. Weiss uses approximations around the point of no evasion to describe the condition under which allowing some degree of evasion can both increase revenue and increase welfare. However, Yitzhaki (1987) shows that, in the examples used by Weiss, the condition that allows successful evasion is identical to the condition that the solution is on the declining portion of the Laffer curve; in this case, *any* reduction of the tax rate would increase welfare and increase revenue. This suggests that the improvement was not caused by allowing evasion. We conclude that neither the practical nor hypothetical relevance of this point has yet been demonstrated.

7.1.3. The optimal extent of enforcement

For a given penalty structure how much resources should be devoted to enforcing the tax laws? Or, in other words, what is the optimal probability of detection, p? Many widely-used textbooks and several IRS commissioners presume that the answer is to increase p until the marginal increase of revenue thus generated equals the marginal resource cost of so doing. As Slemrod and Yitzhaki (1987) show, however, this rule is incorrect. Intuitively, although the cost of increasing p (hiring more auditors, buying better computers, etc.) is a true resource cost, the revenue brought in (through assessed fines as well as higher compliance) does not represent a net gain to the economy, but rather a transfer from private citizens to the government. The correct rule equates the marginal social benefit of reduced evasion to the marginal resource cost; the social cost is not well measured by the increased revenue, but is in this model related to the reduced risk bearing that comes with reduced evasion[42]. This result implies that privatization of revenue collection will inevitably lead to a socially excessive amount of resources devoted to that purpose unless restrictions are put on the resources and behavior of the agency.

7.1.4. Optimal auditing rules

One of the key simplifying assumptions of the Allingham–Sandmo model is that the probability of evasion being detected is fixed and unrelated to any actions of the

[42] Note, though, that Baldry (1984) has shown that complete enforcement of income tax laws (p high enough to deter evasion) is inefficient.

taxpayer. In Section 2.1 we investigated the implication of p increasing with the amount of evasion, but this relationship was exogenously imposed. Other models allow the audit strategy of the tax collection agency (henceforth the IRS) to depend on the report of the taxpayer in a way that maximizes an explicit objective function; the taxpayer, in turn, forms some expectation of what the IRS' auditing rule is, and acts accordingly. In modeling the game between taxpayers and the IRS, researchers have generally assumed that the IRS attempts to maximize net revenue collected. As we discussed earlier, this is not likely to characterize the social-welfare-maximizing solution to how big the enforcement budget ought to be, although it might characterize the optimal allocation of resources for a given IRS budget. Another critical model element is whether it is assumed that the IRS can commit to an announced audit rule, or whether it cannot commit, and therefore will opportunistically audit whatever returns it wishes once the returns are filed. Finally, it is critical whether the IRS budget is assumed to be fixed.

Following Reinganum and Wilde (1985), models of this question generally assume that the probability of audit depends on reported income only. Most papers conclude that the optimal strategy in this context is to randomly audit individuals who report below some threshold level of income. In equilibrium only low-income individuals report honestly, while high-income taxpayers report exactly at the threshold level of income and are never audited. Sanchez and Sobel (1993) derive this result in the context of risk-neutral taxpayers with a continuous distribution of actual income and no labor supply decisions, and where penalties for detected evasion are bounded and exogenously set. Cremer and Gahvari (1996) reach similar conclusions when they allow for endogenous labor supply, although they consider just two types of individuals. Mookherjee and Png (1989) consider risk-averse individuals. Imposing mild restrictions on the level of risk aversion, they show that the optimal policy is characterized by random audits and finite penalties. It is still true that above some income level taxpayers are not audited, but it is no longer true that everyone reporting an income below that level is honest[43]. Scotchmer (1987) relaxes the assumption that the IRS can only observe the taxpayer's report, and instead assumes that it is possible to assign taxpayers to a number of audit classes based on observable characteristics. Although the optimal audit policy within each class is similar to that described above, this policy introduces a regressive bias to the effective tax system, because the agency will audit taxpayers with low-income reports with higher probability than high-report taxpayers, thus making it less attractive for low-income taxpayers to underreport income. This bias may be difficult to undo through the statutory tax system if the tax code cannot depend on the audit class.

This state of affairs provides an obvious temptation to the IRS to reverse its pre-announced audit rule and instead to audit only those taxpayers that report exactly the threshold level of income; those that report below the threshold are, after all, reporting

[43] Note that, in all of the papers in this literature, in the optimal policy taxpayers revealed to be honest are rewarded, a decidedly counterfactual prediction.

truthfully. Because of that temptation, an announced precommitment is not likely to be credible. Describing the equilibrium outcome in the absence of precommitment is more complex, as Andreoni, Erard and Feinstein (1998) discuss. One class of models, first investigated by Graetz, Reinganum and Wilde (1986), introduces a set of taxpayers at each income level who report truthfully regardless of their incentives to do otherwise. This enriches the model because it implies that at each level of income report there are both honest and evading taxpayers. Melumad and Mookherjee (1989) take another tack by assuming that although the government cannot commit to a particular audit policy, it can commit to the total amount spent on audits. In this context they demonstrate that the problem of commitment may be solved by delegating this task to a separate agency, and they describe the optimal contract that guarantees a unique equilibrium and provides incentives for the agency to audit optimally. Such a contract is welfare improving.

In the context of models of tax compliance in which the strategies of both the taxpayers and the IRS are objective-maximizing, the impact of a change in, say, the tax rate, depends on one's forecast of how both sets of actors respond. For example, if the tax rate increases it may become optimal for the IRS to audit more returns; in Graetz, Reinganum and Wilde (1986), with an unconstrained budget, an increase in the tax rate on the high-income taxpayers who are potential evaders decreases evasion. Whether this prediction turns out to be accurate depends on whether in practice the IRS budget increases concomitantly with the tax rate, and there is no empirical evidence that supports this.

7.1.5. Optimal allocation of enforcement resources

Administrative costs are inputs into the revenue raising process. But what should be the target of the administration, and how should economic considerations be introduced into the tax-revenue production function? To address this issue, one has to define the objectives of the tax administration and its production function – how much revenue is produced with different combinations of inputs (subject, of course, to the tax law). Then one can analyze whether the allocation of funds for administration is efficient or to check whether, as Tanzi and Pellechio (1997) put it, "personnel are often assigned to tasks that have low productivity while important functions get unattended".

Yitzhaki and Vakneen (1989) develop a model that introduces microeconomic considerations into the management of tax administration[44]. They assume that the objective of the administration is the maximization of revenue and that taxpayers can be classified into groups based on having returns of similar complexity. These assumptions allow them to present the inspection process of tax returns as a decision tree in which the "inspector" has to spend a given amount of his time to review

[44] See also Wertz (1979), the first modern treatment of the optimal allocation of the work force in a tax collection agency.

the return, and the reaction of the taxpayer (whether to appeal) is determined by the quality of the assessment. If they continue to disagree, the results are determined by the court. The solution to this decision tree problem can be determined in a dynamic programming model. Estimation of the decision tree enables one to estimate the present value of future tax revenue that is collected by each activity of the tax administration. Yitzhaki and Vakneen argue that an administration should equalize the rate of return, in terms of tax revenue, for each activity. This principle should govern sampling of tax returns for inspection, as well as which items on the return to inspect.

7.2. Optimal tax systems

The previous section addressed how to evaluate the appropriate setting of tax enforcement instruments, for a given specification of tax base and rates. The more general problem is to consider all of these aspects simultaneously, what Slemrod (1990) calls the theory of "optimal tax systems". Certainly, the ease of administering various taxes has critical implications for the optimal structure of tax systems. Tax codes which are based on unobservable and practically unmeasurable quantities (such as an ability tax) often look desirable on paper. Integrating the issue of administrative ease into normative tax theory requires a shift of emphasis away from the structure of preferences, which has been the principal focus of optimal tax theory, toward the technology of tax collection.

7.2.1. The choice of tax instruments

With some exceptions, optimal tax theory has dealt with the issue of administering a tax by making extreme assumptions about what kinds of taxes are available to the policymaker. The fundamental results of optimal tax theory depend on implicit assumptions about which taxes can be administered and which cannot. The problem of optimal commodity taxation is interesting only because the possibility of lump-sum taxation is ruled out[45], presumably because it is infeasible. Production efficiency is desirable only if all commodities can be taxed and 100 percent taxation of profits is feasible (or if no profits exist). When consumers are not identical, an ability tax dominates an income tax because it causes no distortion in behavior. The study of optimal income taxation is appropriate when ability taxes are ruled out, usually by appealing to the difficulties of measuring ability for the purpose of basing tax liability on it.

Extreme assumptions about the feasibility of tax instruments are analytically convenient, but incorrect. Ability can be measured, although with some expense and

[45] To be sure, the optimal commodity taxation literature yields insights about the less analytically tractable, but more realistic, multi-person environment. Nevertheless, in most models in which the use of lump-sum taxes is limited, this is done as an assumption rather than as a choice based on the costs and benefits of this instrument.

error. On the other hand, income cannot be measured perfectly, and the degree of accuracy in income measurement depends on the resources expended toward this goal.

Extreme assumptions about the feasibility of tax instruments may also preclude consideration of fundamental changes in policy[46]. For example, a common assumption made in optimal taxation models of developing countries is that income and consumption arising in the agricultural sector are not taxable, although marketable surplus is taxable. Much interesting analysis proceeds from this assumption, but none asks at what point it makes sense for a country to attempt to tax agricultural income, even assuming that it will have only limited success in doing so. There is clear evidence [Riezman and Slemrod (1987)] that countries with low literacy rates tend to rely on highly distorting but (relatively) easily collectable import and export taxes, and shy away from efficient but administratively difficult land taxes. Under what conditions should an imperfect land tax be tried? The answers to these questions depend on the resource cost of administering the new tax instrument relative to its effectiveness, or degree of success. This latter notion has several dimensions, including the true revenue yield and the extent and nature of the mistakes that are made in administration.

Stern (1982) models the choice between an optimal nonlinear income tax, in which income is costlessly observable, and a system of differential lump-sum taxes based on characteristics of taxpayers which can be ascertained with some error. The lump-sum tax system is superior if there are no errors in classifying individuals but, when enough mistakes are made, income taxation may be the preferred system. Stern's analysis recognizes that the two tax systems each have their own information requirements (the lump-sum system requires classifying individuals, the income tax system requires observing incomes). The two systems will also likely have different administrative costs as well, although for the sake of simplicity Stern assumes these costs are identical. Greater accuracy in the classification of individuals could be achieved with higher cost, as could more accurate measurement of income[47].

The optimal tax system framework has also been applied to a more immediately policy-relevant choice, that between direct and indirect taxes. It has frequently been claimed that a shift from income taxation to value added taxation can combat evasion by taxing the spending on goods from the compliant sector by individuals who evade taxes on their income. Boadway, Marchand and Pestieau (1994) consider the optimal mix between a general non-linear income tax and commodity taxes under the assumption that evasion is possible *only* for the income tax. Granting this assumption

[46] The desirability of introducing the choice among tax instruments into the optimal tax problem has been noted by, among others, Hahn (1973) and Atkinson and Stiglitz (1980, p. 363), who state that "for a complete theory of the choice of tax base, a fully articulated model is necessary of the information available to the government and cost of observing the different characteristics". Diamond (1987, p. 640) agrees that this would be ideal, but adds that the standard simplifications "may do little damage to the policy conclusions if the set of feasible policies is well chosen, although the problem of choosing well is a difficult one".

[47] This is an example of the issue of optimal "tagging" discussed in Akerlof (1978).

provides a strong case for commodity taxation to supplement an income tax. The authors recognize that the results would have to be "seriously adjusted" (p. 73, fn. 2) if there is more evasion on indirect than on direct taxes. In contrast, Kesselman (1993) concludes that changing the tax mix toward indirect taxes will have little or none of the claimed anti-evasion effects. Underlying this conclusion is a two-sector model in which the income tax is paid only by workers in the above-ground sector, and the indirect tax is paid completely by above-ground workers but incompletely in the underground sector. This is justified on the grounds that to evade the income tax successfully requires evasion of the indirect tax on output as well, since honest reporting of gross sales for the indirect tax would signal to the authorities the extent of the income tax evasion. Which analysis better captures the reality depends on the technology of tax avoidance and evasion.

7.2.2. Presumptive taxes

The general nature of the optimal tax systems problem is well illustrated by considering a class of taxes – known as presumptive taxes – which are a pervasive element in the tax systems of many developing countries. This kind of tax makes sense in cases where the otherwise desirable tax base is difficult for the tax authorities to measure, verify, and monitor. As a substitute for the desired base is the "presumed" tax base, which is derived from a formula, which itself may be simple or complex, based on more readily monitored items[48]. For example, at one time in Israel a taxi driver had a choice of a tax based on book income or a levy on the accumulated mileage of the taxicab; for shopkeepers, the alternative to a tax on book income was a tax based on the square footage of the shop and other observable characteristics of the business. The wide variety of presumptive taxes used in the developing world is nicely surveyed in Tanzi and Casanegra de Jantscher (1989) and in Rajaraman (1995).

The problem that presumptive taxes address – the difficulty of monitoring certain potential tax bases – is not confined to developing countries, and use of presumptive taxes, albeit with different names, is also widespread in developed countries. Examples include the use of fixed depreciation schedules in place of asset-specific measures of the decline in asset value (economic depreciation), taxation of capital gains on a realization basis, and floors on deductible expenses. Slemrod and Yitzhaki (1994) and Kaplow (1994) analyze the U.S. standard deduction as a presumptive tax; a higher value reduces the administrative and compliance cost of monitoring itemized deductions,

[48] There are two general categories of presumptive taxes. In the first, tax liability is based on an easily monitorable base which is presumably correlated with the ideal tax base. The tax on taxicab mileage or a tax on electricity used by a laundry are examples. In many cases, the monitorable base is a specific input, and the presumptive tax is actually a tax on an input. The second category includes (effective or *de facto*) exemptions or floors, intended to eliminate monitoring costs of "nonfruitful" populations. Examples include exempting businesses with less than a certain number of employees, or floors on deductible expenses.

but it increases horizontal inequity by increasing the range of taxpayers for which the "proper" amount of deduction is replaced by a single number[49].

Upon reflection it is clear that all taxes are presumptive, to some degree. The conceptually pure tax base – be it the flow of income, wealth, sales revenue, or something else – cannot be perfectly measured, and the tax authority is constrained to rely on some correlate of the concept. We label particular taxes as presumptive when the calculation of the tax base deviates in a substantial way from the ideal concept. But there is a pervasive tradeoff between accuracy and the costs of complexity[50].

7.2.3. Optimal commodity taxes

The characterization of optimal commodity taxes is a cornerstone of the standard theory of optimal taxation, dating back to Ramsey (1927). The standard theory, though, assumes that taxes on all commodities can be verified and collected costlessly.

Yitzhaki (1979) investigates the optimal size of the commodity tax base in a representative consumer economy when there is a resource cost, related to administration, to adding goods to the tax base. If, as he assumes, preferences over all goods are Cobb–Douglas, then uniformity of rate for all taxed goods is optimal. Expanding the tax base to cover more goods will reduce the excess burden of taxation, but it increases the administrative cost. The optimal tax system equates the marginal excess burden of raising a dollar of revenue to the marginal administrative cost, and thus minimizes the total resource cost of raising revenue. Wilson (1989) generalizes the framework to constant-elasticity-of-substitution utility functions.

The fact that changes in administrative costs are likely to be discontinuous with respect to changes in tax policy is important in more general treatments of the optimal set of tax instruments. The theory of optimal taxation tells us that, except in special cases, all goods should be taxed at different rates. However, it is likely that administrative cost depends on the number of different tax rates as well as the number of commodities taxed. This is not an issue when one assumes a utility function that implies uniform optimal taxes (e.g., Cobb–Douglas), but is very important under more general preferences; in that case there is a tradeoff between administrative and compliance costs on the one hand and the standard excess burden on the other.

Both the Yitzhaki and Wilson papers assume that a commodity is either in the tax base and taxed at the uniform rate, or out of the tax base entirely. Boadway, Marchand and Pestieau (1994), Cremer and Gahvari (1993) and Kaplow (1990) investigate general characterizations of optimal commodity taxation with evasion, administrative costs and costly enforcement[51]. In Cremer and Gahvari (1993), the optimal tax on a

[49] See also Sadka and Tanzi (1993), who argue in some situations for a presumptive tax on assets as a substitute for a income tax.

[50] Kaplow (1994, 1996) addresses the equity and efficiency issues involved in making this tradeoff.

[51] Skinner and Slemrod (1985) suggest that enforcement policy can be part of an optimal tax system in which the statutory tax rates are constrained to be suboptimal; for example, lax enforcement of a good whose statutory tax rate exceeds the optimal rate may be appropriate.

commodity is, *ceteris paribus,* lower when the elasticity of induced avoidance response to a tax increase is higher; intuitively, this increases the marginal social cost per dollar raised from taxing that commodity.

7.2.4. Optimal progressivity

In the optimal linear income tax literature, where only a demogrant and single marginal tax rate are chosen, what constrains redistribution is the marginal excess burden caused per dollar raised by the marginal tax rate, and the fact that this ratio increases with the marginal tax rate levied. Cremer and Gahvari (1994) investigate how the introduction of evasion and concealment expenses change the optimal setting of a linear income tax, when the audit probability is also optimally chosen. They characterize the optimal marginal tax rate in the presence of evasion, but conclude that one cannot hope for an unambiguous result in general about whether in a model with evasion the marginal tax rate is higher or lower compared to in a model without evasion.

If other aspects of the tax system are not set optimally, there is no presumption that the tax rate that is optimal, given the value of the other instruments, is also the global optimum. To be concrete, if enforcement instruments are set suboptimally, so that the marginal cost of raising revenue is higher than it need be, then the optimal tax rate will appear lower than if the enforcement parameters are set optimally.

The point is that the optimal level of taxes or tax progressivity can be properly assessed only simultaneously with the instruments the government uses to control avoidance and evasion. Slemrod (1994) constructs an example of this issue by modeling a two-person economy in which the only possible response to taxation is avoidance. The government must choose three instruments to maximize social welfare: a demogrant, a (single) marginal tax rate, and an avoidance-control policy denoted p, which at a cost reduces both the level of avoidance and its responsiveness to changes in the marginal tax rate. An example shows that, with p set suboptimally, the optimal policy can be to lower t; however, a superior policy is to raise both p and t. The intuition here is that the calculation of marginal excess burden of the marginal tax rate should be done assuming the other policy instruments are set optimally. Using the metaphor of Okun (1975), the "leak" in the revenue system, which limit both redistribution and the size of the public sector, can be "fixed", albeit at some cost.

Slemrod and Kopczuk (2001) expand on this notion by isolating the effect of a policy instrument on the elasticity of taxable income, which summarizes the magnitude of the behavioral response to taxes that limits optimal progressivity. They formally characterize the optimal elasticity, emphasizing that in many settings it is appropriate to think of this as a policy choice rather than an exogenous constraint. In a special case where the policy instrument is the breadth of the tax base, Slemrod and Kopczuk show that more egalitarian societies will feature lower elasticities of taxable income, as will societies with a lower marginal cost of tax administration. Thus, this research simultaneously addresses optimal progressivity and the optimal ease of collecting taxes, and focuses on a critical difference between real substitution responses on the

one hand and avoidance and evasion responses on the other. Economists nearly always assume that the former is an immutable, or primitive, parameter that is immune to policy (or any kind of) manipulation. Whatever the truth of that assumption as it applies to, say, labor supply response to taxation, it is certainly untenable as it applies to avoidance and evasion responses. Their availability is certainly a (perhaps highly constrained) policy choice. Truly optimal tax policy does not accept the current state of administration and enforcement as given, but instead chooses these aspects and the statutory tax structure together.

7.3. The marginal efficiency cost of funds

A principal theme of this chapter is that acknowledging the range of behavioral responses to taxation suggests a rich set of new empirical and conceptual issues and alters the answers to some fundamental questions of public finance. For some other questions, though, the anatomy of behavioral response may not matter. For example, Feldstein (1999) argues that, for the purpose of calculating the marginal efficiency cost of taxation, the critical parameter is the tax rate elasticity of taxable income, and the etiology of the elasticity – be it increased leisure, evasion, or increased untaxed fringe benefits, for example – is irrelevant. The intuition is that at the margin people are willing to incur a dollar's worth of cost to save a dollar of taxes, and that cost may take the form of a distorted consumption basket, a fee to an accountant, or increased exposure to the risk of punishment for evasion. However, because Feldstein derives this conclusion in a model which allows real substitution response but neither avoidance nor evasion, it begs the question of whether the taxable income elasticity is a sufficient statistic for measuring the efficiency cost of raising taxes and for comparing the relative efficiency of alternative ways to raise revenue.

This problem has been treated in the context of the concept referred to as the marginal cost of funds or marginal efficiency cost of funds, and was developed by Usher (1986), Mayshar (1990, 1991), Wildasin (1984), and Slemrod and Yitzhaki (1996). This model also allows us to place the issues raised above into a more general normative framework. We first discuss the concept in the absence of administrative costs, evasion, or avoidance, and then extend it to apply to these issues[52].

Following Mayshar (1991), assume that the government sets a level of public goods, G, and a vector E of tax policy instruments so as to maximize $V(U^*(E, w), G)$, where w is the wage rate and U^* is the utility derived from private goods. He shows that the optimum is characterized by MBF = MCF_i, where MBF is the social marginal benefit of funds (in terms of private consumption), and MCF_i is the marginal cost of

[52] Although what follows can be generalized to apply to a multi-individual framework and to public goods, here we restrict ourselves to the representative individual model of tax analysis. See Slemrod and Yitzhaki (2001) for a treatment of the more comprehensive problem.

funds of tax instrument i. Mayshar and Yitzhaki (1995) decompose the MCF_i term into:

$$MCF_i = DC_i^* MECF_i, \tag{10}$$

where DC_i is Feldstein's (1972) distributional characteristic of the tax instrument, while $MECF_i$ is the marginal efficiency cost of the tax instrument. In the absence of evasion or avoidance, $MECF_i$ is equal to X_i/MR_i, where X_i is the change in revenue assuming no behavioral response, and MR_i (marginal revenue) allows behavioral response. Thus, in the case of an income tax, X_i/MR_i equals $1/(1 + \varepsilon_i)$, where ε_i is the elasticity of taxable income with respect to tax instrument i.

Note that the above interpretation is not limited to reforms involving tax rates. One may define the marginal cost of funds with respect to marginal changes in any parameter of the tax system (e.g., income brackets, exemption levels, penalties for tax evasion, etc.). Nor does its application rely on an assumption that tax policy has been set optimally. As Slemrod and Yitzhaki (1996) show, away from the optimum the MECF concept can be used to identify incremental changes in the tax system that would increase social welfare.

To see how the MECF can be extended to evasion and avoidance, recall that the potential change in tax revenue (assuming an inelastic base) is X_i but, because of taxpayers' response, the government collects only MR_i. We can divide the potential tax X_i into two components as follows:

$$X_i = (X_i - MR_i) + MR_i, \tag{11}$$

where MR_i dollars are collected and $(X_i - MR_i)$ "leaks" outside the tax system. The critical question is how to evaluate, from a social point of view, the leaked dollars. To do this one must ask how much a taxpayer is ready to expend (on the margin) to save a dollar of taxes or, alternatively, how much utility loss he is willing to suffer to save a dollar of taxes. The answer is that a rational taxpayer will be ready to sacrifice up to, but no more than, one dollar in order to save a dollar of taxes. Hence, on the margin the private cost, which is equal to "leaked" dollars multiplied by their cost per dollar, is $X_i - MR_i$; the collection of MR_i dollars results in a loss of $(X_i - MR_i)$ to the taxpayer over and above the taxes paid. If we assume that the utility loss to the individual (private cost) of the leaked tax revenue should be accorded the same social cost as the utility loss due to the taxes paid, then the cost to society of transferring a dollar to the government is $(X_i - MR_i)/MR_i = (X_i/MR_i) - 1$. The total marginal cost to the individual taxpayer, including the taxes paid, is X_i/MR_i.

Consider now a taxpayer who also has the option to evade part of the additional tax. On the margin, he would be ready to sacrifice utility valued at one dollar (in additional risk bearing due to evasion and/or due to substitution to cheaper but less rewarding activities) in order to save a dollar of taxes. Hence, we do not have to know whether the "leak" was through evasion or real substitution to evaluate the costs to society.

The same rule applies to avoidance activity and, in fact, to any activity under taxpayer control. Therefore, all one needs to know is the potential tax (i.e., assuming an inelastic tax base) that will be collected from a change of a parameter of the tax system, and the actual change (taking into account all behavioral responses) in order to evaluate the marginal efficiency cost of raising revenue. It is in this sense that Feldstein's (1999) claim about the central importance of the elasticity of taxable income generalizes to avoidance and evasion.

Calculating the MECF involves two critical assumptions that deserve further attention. The first of these is that at the margin the taxpayer sacrifices exactly one dollar (instead of *up to* one dollar) to reduce tax liability by one dollar. However, it may be that the taxpayer is at a corner solution with respect to behavioral response, so that the marginal utility loss may be less than a dollar. For an example of a taxpayer at a corner, consider the case of Individual Retirement Accounts (IRAs). An employee can contribute up to $2000 per year into an IRA, deduct the contribution from taxable income, pay no tax on accrued earnings in the account, and pay tax on the principal when withdrawn. Although IRAs were designed to increase saving, there is nothing to prevent an individual who in the absence of taxes would have invested $4000 in a similar account from diverting $2000 into the IRA. There may be a cost to this, as IRAs have early withdrawal penalties which in some cases limit the flexibility of using these funds. Thus, contributing to an IRA can save taxes, does not require a change in one's consumption basket, but may entail some cost. However, it cannot be presumed that, at the margin of an IRA contribution, the private value of the sacrifice is equal to the tax saving; the IRA contribution is limited to $2000 only because of the statutory limit on contributions. As another example, consider the MECF of raising the tax rate on labor income in a situation where, in an economy with two taxpayers, one taxpayer reports no labor income at all and, at that corner, is bearing risk valued at 20 cents (rather than a dollar, as would be true at an interior solution) to evade, including penalties, an expected value of one dollar. The other taxpayer, with identical labor income, reports all of it. Assuming no labor supply or avoidance response, the MECF with respect to an increased tax rate is 1.2.

To take account of the possibility of the taxpayer being at such corner solutions, one can generalize the expression for the MECF by introducing a parameter γ, $0 < \gamma \leqslant 1$, which is a weighted average of the marginal value to the taxpayers of the leaked revenue, $X_i - \mathrm{MR}_i$. Introducing γ reduces the simplicity of the MECF expression because its value varies depending on the situation under study.

The second critical assumption is that the cost borne by taxpayers in the process of reducing tax liability is equivalent to the social cost. This is certainly true in many situations, such as when the private cost takes the form of a distorted consumption basket. But in some cases the private cost is not identical to the social cost. An example is when the act of the taxpayer causes some externality. Consider the case where being caught evading imposes a stigma on the taxpayer, as in Benjamini and Maital (1985) or Gordon (1989), and assume that the larger the number of evaders the lower the stigma attached to each act. In this case the social cost of evading taxes diverges from

the private cost because the potential evader does not take into account the impact of his action on other members of the society.

Fines for tax evasion present another example of the potential divergence between the private and social costs of tax-reducing activities. The possibility of a fine for detected tax evasion is certainly viewed as a cost by the taxpayer, but from society's point of view it reduces the amount of revenue that would otherwise have to be collected. (This is in contrast to imprisonment, unless the prisoner is forced to produce socially valuable products while imprisoned.) Thus, the MR term should include fine collections. Note that, if the fine itself is the policy instrument, this argument implies that its MECF could be close to zero, and almost certainly less than one, making an increase in fines look like an attractive policy option indeed. As discussed in Section 7.1.1, there are reasons unrelated to efficiency cost minimization which render undesirable increasing fines for tax evasion without limit.

Applying the MECF rule to administrative and compliance issues clarifies the common thread running through models of optimal tax systems. In the generic problem, there are two ways to raise revenue: to increase a set of tax rates, and by so doing to increase excess burden, or via an alternative which involves increasing administrative costs [e.g., by broadening the tax base as in Yitzhaki (1979), or by increasing the probability of a tax audit, as in Slemrod and Yitzhaki (1987)]. On the margin, it is optimal to equalize the marginal costs of raising revenue under the two alternatives. If one defines the costs of taxation as deadweight loss plus administrative costs, at an optimum the MECF of each tax rate should be equal to the MECF of administrative improvements that raise revenue. In calculating the MECF of administrative improvements, it is important to account for the fact that these expenses come out of funds that were presumably raised with tax instruments that have an MECF in excess of one. In other words, administrative improvements that raise net revenue decrease the excess burden; hence, on the margin and for given revenue, the saving in excess burden should be equal to the increase in administrative costs. In this way, the MECF criterion can be applied to tax administration, too[53].

Compliance costs are additional costs imposed on the taxpayer. Therefore, they should be added to the burden imposed on the taxpayer. They serve as a substitute to administrative costs, but the expenses are borne directly by the taxpayer rather than through the government budget.

The revised MECF that includes all these factors, derived and discussed in Slemrod and Yitzhaki (1996), is

$$\mathrm{MECF}_i = \frac{\gamma(X_i - \mathrm{MR}_i) + C_i + \mathrm{MR}_i}{\mathrm{MR}_i - A_i}, \tag{12}$$

[53] Yitzhaki and Vakneen (1989) use the term "the shadow price of a tax inspector", which is the revenue collected by adding another tax inspector. Note that the MECF is actually the reciprocal of the shadow price of a tax inspector.

where γ is the social value of the utility the taxpayer is sacrificing at the margin in order to save a dollar of tax. C_i is the marginal private compliance cost associated with the ith instrument, A_i is the marginal administrative cost, and $MR_i - A_i$ is the net revenue collected at the margin. The intuitive interpretation of the expression is the same as before, with some qualifications. The potential tax is X_i. $X_i - MR_i$ is leaked at a social cost of γ per dollar, MR_i is collected by the government, and C_i is the additional involuntary compliance cost. Hence, the total burden on society is the sum of those components. Of the MR_i collected by the government, A_i is spent on administration, leaving $MR_i - A_i$ in the coffers. The MECF is the burden on society divided by what is collected after subtracting the cost of doing business. This yields the marginal costs of a dollar collected.

Because in Equation (12), C_i is added in the numerator and A_i is subtracted in the denominator, the key conceptual difference between the two is explicit – only the latter uses revenue raised from taxpayers. To illustrate this difference, consider that a tax for which $C_i = MR_i$ (with A_i and $X_i - MR_i = 0$) might conceivably be part of an optimal tax regime (if the MECFs of other instruments exceed two), but it would never be optimal to have $A_i = MR_i$, for at the margin this instrument has social cost but raises no revenue.

As emphasized in Slemrod (1998), applying this notion using empirical estimates of the taxable income elasticity must be done with care. Foremost is the need to consider the elasticity of the *present value* of tax revenues. Recall that a class of avoidance responses involves the retiming of taxable-income-generating events. If a tax policy change causes retiming, focusing only on the revenues in a subset of periods will bias the findings. For example, the taxable income response to an anticipated future decrease in tax rates must consider the lost revenue in the period before the tax rate changes[54]. Similarly, if a tax change causes an increase in deferred compensation, the increased future tax liability must be netted against any decline in current tax payments. Furthermore, any change in taxable income in one tax base must be netted against changes in taxable income in other bases. For example, if a decline in personal tax rates causes a shift from C corporation status to S corporation status, the increased personal taxable income must be netted against decreases in corporate taxable income.

8. Conclusion

The possibilities for evasion and the difficulties of administration have always shaped tax systems. Until recently, formal analysis of taxation largely ignored these realities. After a quarter of a century of research on the topic, it is time to put to rest the claim

[54] Feldstein (1999) recognizes the timing effect, but whether the empirical analysis of the 1986 tax changes from which he derives a taxable income elasticity [Feldstein (1995)] does or does not is a controversial question, with Slemrod (1996b) arguing that timing effects dominate the results.

that this is an understudied area. Instead, it is a vibrant area of research that has clarified the positive and normative analysis of taxation.

The research has clarified that when the tax structure changes, people may alter their consumption basket, but they also may call and give new instructions to their accountant, change their reports to the IRS, change the timing of transactions, and effect a set of other actions that do not directly involve a change in their consumption basket. In many cases, particularly for high-income taxpayers, this latter set of responses has larger revenue and welfare implications than the real substitution responses, such as labor supply, that tax analysis has traditionally focused on.

Early models of this area focused on tax evasion, modeled as a gamble against the enforcement capability of the state. More recently, the literature has examined more general models of the technology of avoidance, with the additional risk bearing caused by tax evasion either being a special case of this technology or one aspect of the cost of changing behavior to reduce tax liability. A critical aspect of this technology is whether the avoidance is inframarginal, in which case only income effects are involved, or whether its cost depends on other aspects of behavior. If the latter is true, the choice of consumption basket and avoidance become intertwined because certain activities may facilitate avoidance, which alters their effective relative price or return.

Acknowledging the variety of behavioral responses to taxation greatly enriches the normative analysis of taxation. It changes the answers to traditional subjects of inquiry, such as incidence, optimal progressivity, optimal commodity taxation, and the optimal mix between income and consumption taxes. It also raises a whole new set of policy questions, such as the appropriate level of resources to devote to administration and enforcement, and how those resources should be deployed. A recurring question that runs throughout this chapter is whether the standard toolkit of positive and empirical analysis can be applied to avoidance, evasion, and administration. The answer is a qualified yes, as this chapter hopefully demonstrates.

In one respect, though, the policy perspective does change in an important way. The magnitude of real substitution response, such as labor supply, to taxes is presumed to be an immutable function of preferences, and not susceptible to policy manipulation in a free society. With respect to avoidance and evasion, though, this hands-off approach is not appropriate. On the contrary, there are a variety of policy instruments that can affect the magnitude and nature of avoidance and evasion response, ranging from the activities of the enforcement agency to how tightly drawn are rules and regulations. The same kind of cost–benefit calculus applies to the choice of these instruments, implying that the elasticity of behavioral response is itself a policy instrument, to be chosen optimally.

A key challenge for the future is to add more empirical content to the theoretical models of taxpayer and tax agency behavior. This will require, *inter alia,* addressing the technology of raising and avoiding taxes. This is the analogue to the critical role for traditional taxation theory of the empirical investigation of the structure of individuals' preferences. Although by their nature the appropriate data are often difficult to come

by, new approaches such as controlled field experiments and analysis of changes in tax administration are promising.

It would also be fruitful to incorporate public choice considerations into the analysis. In some case administrative difficulties as well as widespread avoidance and evasion are caused by the inability of compromise-seeking legislators to agree upon a well-defined law. Furthermore, there is apparently no political constituency for tax simplicity and facilitated adminstration. Combining analysis of the public choice mechanisms that produce tax systems with the kind of normative analyses discussed in this chapter may lead to a more complete understanding of the reality of taxation.

References

Agell, J., and M. Persson (1990), "Tax arbitrage and the redistributive properties of progressive income taxation", Economic Letters 34(4):357–361.

Akerlof, G. (1978), "The economics of 'tagging' as applied to the optimal income tax, welfare programs, and manpower planning", American Economic Review 68(1):8–19.

Allingham, M.G., and A. Sandmo (1972), "Income tax evasion: a theoretical analysis", Journal of Public Economics 1(3–4):323–338.

Alm, J. (1988), "Compliance costs and the tax avoidance-tax evasion decision", Public Finance Quarterly 16(1):31–66.

Alm, J. (1999), "Tax compliance and administration", in: W. Bartley Hildreth and James A. Richardson, eds., Handbook on Taxation (Marcel Dekker, New York) pp. 741–768.

Alm, J., R. Bahl and M.N. Murray (1991), "Tax base erosion in developing countries", Economic Development and Cultural Change 39(4):849–872.

Alm, J., B.R. Jackson and M. McKee (1992), "Estimating the determinants of taxpayer compliance with experimental data", National Tax Journal 45(1):107–114.

Alt, J. (1983), "The evolution of tax structures", Public Choice 41(1):181–222.

Andreoni, J. (1992), "IRS as loan shark: tax compliance with borrowing constraints", Journal of Public Economics 49(1):35–46.

Andreoni, J., B. Erard and J. Feinstein (1998), "Tax compliance", Journal of Economic Literature 36(2):818–860.

Atkinson, A.B., and J.E. Stiglitz (1980), Lectures in Public Economics (McGraw-Hill, New York).

Baldry, J.C. (1979), "Tax evasion and labor supply", Economics Letters 3(1):53–56.

Baldry, J.C. (1984), "The enforcement of income tax laws: efficiency implications", Economic Record 60(169):156–159.

Baldry, J.C. (1987), "Income tax evasion and the tax schedule: some experimental results", Public Finance 42(3):357–383.

Balke, N.S., and G.W. Gardner (1991), "Tax collection costs and the size of government", Mimeo (Southern Methodist University).

Beck, P., and W.O. Jung (1987), "An economic model of taxpayer compliance under complexity and uncertainty", Journal of Accounting and Public Policy 8:1–27.

Becker, G.S. (1968), "Crime and punishment: an economic approach", Journal of Political Economy 76(2):169–217.

Benjamini, Y., and S. Maital (1985), "Optimal tax evasion and optimal tax policy", in: W. Gartner and A. Wenig, eds., The Economics of the Shadow Economy (Springer, Berlin) pp. 245–264.

Beron, K.J., H.V. Tauchen and A.D. Witte (1992), "The effect of audits and socioeconomic variables on compliance", in: Joel Slemrod, ed., Why People Pay Taxes (University of Michigan Press, Ann Arbor) pp. 67–89.

Bird, R.M. (1983), "Income tax reform in developing countries; the administrative dimension", Bulletin for International Fiscal Documentation 37:3–14.
Bishop, J.A., K.V. Chow, J.P. Formby and C.-C. Ho (1994), "The redistributive effects of noncompliance and tax evasion in the U.S", in: John Creedy, ed., Taxation, Poverty, and Income Distribution (Edward Elgar, Aldershot) pp. 28–47.
Blough, R. (1952), The Federal Taxing Process (Prentice-Hall, Washington, D.C.)
Blumenthal, M., C.W. Christian and J. Slemrod (2001), "Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota", National Tax Journal 54(1):125–138.
Boadway, R., M. Marchand and P.M. Pestieau (1994), "Towards a theory of the direct-indirect tax mix", Journal of Public Economics 55(1):71–88.
Burman, L., and W. Randolph (1994), "Measuring permanent responses to capital gains taxes in panel data", American Economic Review 84(4):794–809.
Burman, L., K. Clausing and J. O'Hare (1994), "Tax reform and realizations of capital gains in 1986", National Tax Journal 45(1):1–18.
Casanegra de Jantscher, M. (1990), "Administering the VAT", in: Malcolm Gillis, Carl S. Shoup and Gerardo Sicat, eds., Value-Added Taxation in Developing Countries (The World Bank, Washington, D.C.) pp. 171–179.
Christian, C.W. (1994), "Voluntary compliance with the individual income tax: results from the 1988 TCMP study", in: The IRS Research Bulletin, 1993/1994, Publication 1500 (Rev. 9-94) (Internal Revenue Service, Washington, D.C.) pp. 35–42.
Clotfelter, C.T. (1983), "Tax evasion and tax rates: an analysis of individual returns", Review of Economics and Statistics 65(3):363–373.
Cowell, F.A. (1990a), "Tax sheltering and the cost of evasion", Oxford Economic Papers 42(1):231–243.
Cowell, F.A. (1990b), Cheating the Government: The Economics of Evasion (MIT Press, Cambridge).
Cremer, H., and F. Gahvari (1993), "Tax evasion and optimal commodity taxation", Journal of Public Economics 50(2):261–275.
Cremer, H., and F. Gahvari (1994), "Tax evasion, concealment, and the optimal linear income tax", Scandinavian Journal of Economics 96(2):219–239.
Cremer, H., and F. Gahvari (1996), "Tax evasion and the optimum general income tax", Journal of Public Economics 60(2):235–249.
Cross, R., and G.K. Shaw (1982), "On the economics of tax aversion", Public Finance 37(1):36–47.
Diamond, P.A. (1987), "Optimal tax theory and development policy: directions for future research", in: David Newbery and Nicholas H. Stern, eds., The Theory of Taxation for Developing Countries (Oxford University Press, New York) pp. 638–647.
Dixit, A.K. (1991), "The optimal mix of inflationary finance and commodity taxation with collection lags", International Monetary Fund Staff Papers 38(3):643–654.
Domar, E.D., and R.A. Musgrave (1944), "Proportional income taxation and risk-taking", Quarterly Journal of Economics 58:388–423.
Dubin, J.A., and L.L. Wilde (1988), "The empirical analysis of federal income tax auditing and compliance", National Tax Journal 41(1):61–74.
Dubin, J.A., M.J. Graetz and L.L. Wilde (1990), "The effect of audit rates on the federal individual income tax, 1977–1986", National Tax Journal 43(4):395–409.
Engel, E., and J.R. Hines Jr (1999), "Understanding tax evasion dynamics", Working Paper 6903 (National Bureau of Economic Research).
Erard, B. (1997), "Self-selection with measurement errors: a microeconometric analysis of the decision to seek tax assistance and its implications for tax compliance", Journal of Econometrics 81(2):319–356.
Feige, E.L. (1979), "How big is the irregular economy?" Challenge 22(5):5–13.
Feinstein, J. (1991), "An econometric analysis of income tax evasion and its detection", RAND Journal of Economics 22(1):14–35.
Feldstein, M. (1972), "Distributional equity and the optimal structure of public prices", American Economic Review 62(1):32–36.

Feldstein, M. (1995), "The effect of marginal tax rates on taxable income: a panel study of the 1986 Tax Reform Act", Journal of Political Economy 103(3):551–572.
Feldstein, M. (1999), "Tax avoidance and the deadweight loss of the income tax", Review of Economics and Statistics 81(4):674–680.
Goolsbee, A. (2000), "What happens when you tax the rich: evidence from executive compensation", Journal of Political Economy 108(2):352–378.
Gordon, J. (1989), "Individual morality and reputation costs as deterrents to tax evasion", European Economic Review 33(4):797–805.
Gordon, R.H., and J.K. MacKie-Mason (1990), "Effects of the Tax Reform Act of 1986 on corporate financial policy and organizational form", in: Joel Slemrod, ed., Do Taxes Matter? The Impact of the Tax Reform Act of 1986 (MIT Press, Cambridge, MA) pp. 91–131.
Gordon, R.H., and J.K. MacKie-Mason (1997), "How much do high taxes discourage incorporation?" Journal of Finance 52(2):477–505.
Gordon, R.H., and J. Slemrod (1988), "Do we collect any revenue from taxing capital income?" in: Lawrence Summers, ed., Tax Policy and the Economy (MIT Press and National Bureau of Economic Research, Cambridge, MA) pp. 89–130.
Gordon, R.H., and J. Slemrod (2000), "Are 'real' responses to taxation simply shifting between corporate and personal tax bases?" in: Joel Slemrod, ed., Does Atlas Shrug? The Economic Consequences of Taxing the Rich (Harvard University Press, Cambridge; the Russell Sage Foundation, New York) pp. 240–280.
Gordon, R.H., and J.D. Wilson (1989), "Measuring the efficiency cost of taxing risky capital income", American Economic Review 79(3):427–439.
Graetz, M.J., J.F. Reinganum and L.L. Wilde (1986), "The tax compliance game: toward an interactive theory of tax enforcement", Journal of Law, Economics, and Organization 2(1):1–32.
Groves, H.M. (1974), Tax Philosophers: Two Hundred Years of Thought in Great Britain and the United States (The University of Wisconsin Press, Madison). Published posthumously; edited by Donald J. Curran.
Gutmann, P.M. (1977), "The subterranean economy", Financial Analysts Journal 33(6):26–34.
Hahn, F. (1973), "On optimum taxation", Journal of Economic Theory 6(1):96–106.
Hinrichs, H.H. (1966), A General Theory of Tax Structure Change During Economic Development (The Law School of Harvard University, Cambridge, MA).
Internal Revenue Service (1996), "Federal tax compliance research: individual income tax gap estimates for 1985, 1988, and 1992", Publication 1415 (Internal Revenue Service, Washington, D.C.)
Kanbur, S.M. (1979), "On risk taking and the personal distribution of income", Journal of Political Economy 87(4):769–797.
Kaplow, L. (1990), "Optimal taxation with costly enforcement and evasion", Journal of Public Economics 43(2):221–236.
Kaplow, L. (1994), "The standard deduction and floors in the income tax", Tax Law Review 50:1–31.
Kaplow, L. (1996), "How tax complexity and enforcement affect the equity and efficiency of the income tax", National Tax Journal 45(1):135–150.
Kau, J.B., and P.H. Rubin (1981), "The size of government", Public Choice 37(2):261–274.
Kay, J.A. (1980), "The anatomy of tax avoidance", in: Income Distribution: The Limits to Redistribution; Proceedings of the 31st Symposium of the Colston Research Society, University of Bristol (John Wright) pp. 135–148.
Kesselman, J.R. (1989), "Income tax evasion: an intersectoral analysis" Journal of Public Economics 38(2):137–182.
Kesselman, J.R. (1993), "Evasion effects of changing the tax mix", The Economic Record 69(205): 131–148.
Klepper, S., and D. Nagin (1989), "The anatomy of tax evasion", Journal of Law, Economics, and Organization 5(1):1–24.
Kolm, S.-C. (1973), "A note on optimum tax evasion", Journal of Public Economics 2(3):265–270.

Long, S. (1981), "Social control in the civil law: the case of income tax enforcement", in: H.L. Ross, ed., Law and Deviance (Sage Publications, Beverly Hills) pp. 185–214.
Maki, D.M. (1996), "Portfolio shuffling and tax reform", National Tax Journal 49(3):317–329.
Mansfield, C. (1988), "Tax administration in developing countries", International Monetary Fund Staff Papers 35(1):181–197.
Mayshar, J. (1990), "On measures of excess burden and their application", Journal of Public Economics 43(3):263–289.
Mayshar, J. (1991), "Taxation with costly administration", Scandinavian Journal of Economics 93(1): 75–88.
Mayshar, J., and S. Yitzhaki (1995), "Dalton-improving tax reform", American Economic Review 85(4):793–807.
Melumad, N., and D. Mookherjee (1989), "Delegation as commitment: the case of income tax audits", RAND Journal of Economics 104(2):139–163.
Mookherjee, D., and I.P.L. Png (1989), "Optimal auditing, insurance, and redistribution", Quarterly Journal of Economics 104(2):399–415.
Musgrave, R.A. (1969), Fiscal Systems (Yale University Press, New Haven and London).
Okun, A.M. (1975), Equality and Efficiency: The Big Tradeoff (The Brookings Institution, Washington, D.C.).
Pencavel, J. (1979), "A note on income tax evasion, labor supply and nonlinear tax schedules", Journal of Public Economics 12(1):115–124.
Polinsky, A.M., and S. Shavell (2000), "The economic theory of public enforcement of law", Journal of Economic Literature 38(1):45–76.
Rajaraman, I. (1995), "Presumptive direct taxation: lessons from experience in developing countries", Economic and Political Weekly (Mombai, India), May 6-13.
Ramsey, F.P. (1927), "A contribution to the theory of taxation", Economic Journal 37:47–61.
Randolph, W. (1995), "Dynamic income, progressive taxes, and the timing of charitable contributions", Journal of Political Economy 103(4):70–138.
Reinganum, J.F., and L.L. Wilde (1985), "Income tax compliance in a principal agent framework", Journal of Public Economics 26(1):1–18.
Richupan, S. (1984), "Measuring tax evasion", Finance and Development 21(4):38–40.
Riezman, R., and J. Slemrod (1987), "Tariffs and collection costs", Weltwirtschaftliches Archiv 123(3): 545–549.
Rosen, H.S. (1976), "Tax illusion and the labor supply of married women", Review of Economics and Statistics 58(2):485–507.
Roth, J.A., J.T. Scholz and A.D. Witte (1989), Taxpayer Compliance, Vol. 1, An Agenda for Research; Vol. 2, Social Science Perspectives (University of Pennsylvania Press, Philadelphia).
Sadka, E., and V. Tanzi (1993), "A tax on gross assets of enterprises as a form of presumptive taxation", Bulletin for International Fiscal Documentation 47(2):66–73.
Sanchez, I., and J. Sobel (1993), "Hierarchical design and enforcement of income tax policies", Journal of Public Economics 50(3):345–369.
Sandford, C. (1973), The Hidden Costs of Taxation (Institute for Fiscal Studies, London).
Sandford, C., ed. (1995), Tax Compliance Costs: Measurement and Policy (Fiscal Publications, Bath).
Scholz, J.K. (1994), "Tax progressivity and household portfolios: descriptive evidence from the surveys of consumer finances", in: Joel Slemrod, ed., Tax Progressivity and Income Inequality (Cambridge University Press, Cambridge) pp. 219–267.
Scotchmer, S. (1987), "Audit classes and tax enforcement policy", American Economic Review 77(2): 229–233.
Scotchmer, S. (1989), "Who profits from taxpayer confusion?" Economics Letters 29(1):49–55.
Scotchmer, S., and J. Slemrod (1989), "Randomness in tax enforcement", Journal of Public Economics 38(1):17–32.

Shoup, C., R. Blough and M. Newcomer, eds (1937), Facing the Tax Problem (Twentieth Century Fund, New York).
Skinner, J., and J. Slemrod (1985), "An economic perspective on tax evasion", National Tax Journal 38(3):345–353.
Slemrod, J. (1990), "Optimal taxation and optimal tax systems", Journal of Economic Perspectives 4(1):157–178.
Slemrod, J. (1994), "Fixing the leak in Okun's bucket: optimal tax progressivity when avoidance can be controlled", Journal of Public Economics 55(1):41–51.
Slemrod, J. (1996a), "Which is the simplest tax system of them all?" in: Henry Aaron and William Gale, eds., Economic Effects of Fundamental Tax Reform (The Brookings Institution, Washington, D.C.) pp. 355–391.
Slemrod, J. (1996b), "High-income families and the tax changes of the 1980s: the anatomy of behavioral response", in: Martin Feldstein and James Poterba, eds., Empirical Foundations of Household Taxation (University of Chicago Press, Chicago; and NBER) pp. 169–192.
Slemrod, J. (1998), "Methodological issues in measuring and interpreting taxable income elasticities", National Tax Journal 51(4):773–788.
Slemrod, J. (2001), "A general model of the behavioral response to taxation", International Tax and Public Finance 8(2):119–128.
Slemrod, J., and W. Kopczuk (2001), "The optimal elasticity of taxable income", Journal of Public Economics, forthcoming.
Slemrod, J., and S. Yitzhaki (1987), "The optimal size of a tax collection agency", Scandinavian Journal of Economics 89(2):183–192.
Slemrod, J., and S. Yitzhaki (1994), "Analyzing the standard deduction as a presumptive tax", International Tax and Public Finance 1(1):25–34.
Slemrod, J., and S. Yitzhaki (1996), "The cost of taxation and the marginal efficiency cost of funds", International Monetary Fund Staff Papers 43(1):172–198.
Slemrod, J., and S. Yitzhaki (2001), "Integrating expenditure and tax decisions: the marginal cost of funds and the marginal benefit of projects", National Tax Journal 54(2):189–201.
Slemrod, J., M. Blumenthal and C.W. Christian (2001), "Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota", Journal of Public Economics 79(3): 455–483.
Sørensen, P.B. (1994), "From the global income tax to the dual income tax: recent tax reforms in the Nordic countries", International Tax and Public Finance 1(1):57–79.
Stern, N.H. (1982), "Optimum taxation with errors in administration", Journal of Public Economics 17(2):181–211.
Steuerle, C.E. (1985), Taxes, Loans, and Inflation: How the Nation's Wealth Becomes Misallocated (The Brookings Institution, Washington, D.C.).
Stiglitz, J.E. (1982), "Utilitarianism and horizontal equity: the case for random taxes", Journal of Public Economics 18(1):1–33.
Stiglitz, J.E. (1985), "The general theory of tax avoidance", National Tax Journal 38(3):325–337.
Tanzi, V. (1980), "The underground economy in the United States: estimates and implications", Banco Nazionale del Lavoro Quarterly Review 135:427–453.
Tanzi, V. (1992), "Theory and policy: a comment on Dixit and on current theory", International Monetary Fund Staff Papers 39(4):957–966.
Tanzi, V., and M. Casanegra de Jantscher (1989), "The use of presumptive income in modern tax systems", in: Aldo Chiancone and Ken Messere, eds., Changes in Revenue Structures; Proceedings of the 42nd Congress of the International Institute of Public Finance (Wayne University Press, Detroit) pp. 37–51.
Tanzi, V., and A. Pellechio (1997), "The reform of tax administration", in: Christopher Clague, ed., Institutions and Economic Development: Growth and Governance in Less-Developed and Post-Socialist Countries (Johns Hopkins Press, Baltimore) pp. 273–292.

Usher, D. (1986), "Tax evasion and the marginal cost of public funds", Economic Inquiry 24(4):563–586.
Weiss, L. (1976), "The desirability of cheating incentives and randomness in the optimal income tax", Journal of Political Economy 84(6):1343–1352.
Wertz, K.L. (1979), "Allocation by and output of a tax-administering agency", National Tax Journal 32(2):143–156.
Wildasin, D.E. (1984), "On public good provision with distortionary taxation", Economic Inquiry 22(2):227–243.
Wilson, J.D. (1989), "On the optimal tax base for commodity taxation", American Economic Review 79(5):1196–1206.
Yaniv, G. (1988), "Withholding and non-withheld tax evasion", Journal of Public Economics 35(2): 183–204.
Yaniv, G. (1992), "Collaborated employee–employer tax evasion", Public Finance 47(2):312–321.
Yitzhaki, S. (1974), "A note on 'income tax evasion: a theoretical analysis", Journal of Public Economics 3(2):201–202.
Yitzhaki, S. (1979), "A note on optimal taxation and administrative costs", American Economic Review 69(2):475–480.
Yitzhaki, S. (1987), "On the excess burden of tax evasion", Public Finance Quarterly 15(2):123–137.
Yitzhaki, S., and Y. Vakneen (1989), "On the shadow price of a tax inspector", Public Finance 44(3):492–505.

Chapter 23

ENVIRONMENTAL TAXATION AND REGULATION*

A. LANS BOVENBERG

CentER, Tilburg University, Netherlands

LAWRENCE H. GOULDER

Stanford University, Resources for the Future, and NBER

Contents

* The authors are grateful to Alan Auerbach, Dallas Burtraw, Louis Kaplow, Ian Parry, Steven Shavell, and Robert Stavins for helpful suggestions; to Koshy Mathai, Jeffrey Muller, and Roberton Williams III for excellent research assistance; and to the National Science Foundation (Grant SBR-9310362) and US Environmental Protection Agency (Grant R825313-01) for financial support.

Handbook of Public Economics, Volume 3, Edited by A.J. Auerbach and M. Feldstein

Abstract

This chapter examines government policy alternatives for protecting the environment. We compare environmentally motivated taxes and various non-tax environmental policy instruments in terms of their efficiency and distributional impacts. Much of the analysis is performed in a second-best setting where the government relies on distortionary taxes to finance some of its budget. The chapter indicates that in this setting, general-equilibrium considerations have first-order importance in the evaluation of environmental policies. Indeed, some of the most important impacts of environmental policies take place outside of the market that is targeted for regulation.

Section 2 examines the optimal level of environmental taxes, both in the absence of other taxes and in the second-best setting. Section 3 analyzes the impacts of environmental tax reforms, concentrating on revenue-neutral policies in which revenues from environmental taxes are used to finance cuts in ordinary, distortionary taxes. Here we explore in particular the circumstances under which the "recycling" of revenues from environmental taxes through cuts in distortionary taxes can eliminate the non-environmental costs of such reforms – an issue that has sparked considerable interest in recent years. Section 4 compares environmental taxes with other policy instruments – including emissions quotas, performance standards, and subsidies to abatement – in economies with pre-existing distortionary taxes. We first compare these instruments assuming that policymakers face no uncertainties as to firms' abatement costs or the benefits of environmental improvement, and then expand the analysis to explore how uncertainty on the part of regulators and the associated monitoring and enforcement costs affect the choice among alternative policy instruments. Section 5 concentrates on the trade-offs between efficiency and distribution in a second-best setting. Section 6 offers conclusions.

Keywords

second best, general equilibrium, marginal cost of public funds, Pigouvian rule, efficiency, distribution, uncertainty, employment, environmental taxation, environmental regulation, emission permits, performance standards

JEL classification: D5, D6, H2, H4, Q2, Q4

1. Introduction

Many aspects of the natural environment are public goods. Air and water quality are shared (nonrival) goods, as are the wildlife and natural landmarks enjoyed in forests and wilderness areas. Property rights for environmental resources are often difficult, if not impossible, to assign; hence private ownership is the exception rather than the rule. The absence of private ownership implies a lack of markets for important environmental amenities. Since no one owns the air, for example, no one can charge for use of the air (that is, for degradation of the air from pollution) and thus no market arises for air quality. The absence of markets, in turn, implies inefficient use of the environment. Without government intervention, decentralized market economies tend to generate an inefficient balance between the "supply" of environmental goods and services (that is, the levels of environmental quality) and the supply of other goods and services.

Inefficient market outcomes suggest a role for the public sector. In principle, government environmental policies can provide the missing markets or improve the functioning of existing ones. In practice, however, government agencies face daunting challenges in the environmental arena. It is exceptionally difficult to determine, or even approximate, the efficient degree of environmental protection. This requires information about the value that the public attaches to environmental improvement. Because of the public-goods nature of environmental amenities, such values are not easily identified. Moreover, determining the appropriate form of government intervention is a difficult enterprise. Pigou's classic contribution showed that taxes could be employed to account for environmental externalities. However, in realistic policy settings, where other (non-environmental) distortions are present, where information about benefits and costs is incomplete, and where distributional concerns and political constraints must be considered along with the efficiency outcomes, the choice among instruments for environmental protection becomes more complicated. In these circumstances it may no longer be optimal to introduce taxes along Pigouvian lines – or even to introduce taxes at all – in order to protect the environment.

In this chapter we examine government policy alternatives for protecting the environment. We pay considerable attention to how taxes can be employed to achieve this goal. However, we will also examine alternatives to taxes, such as emissions quotas and performance standards, and compare these alternatives with taxes along efficiency and other dimensions.

Traditionally, analyses of environmental taxes and other regulations have been partial-equilibrium in nature. Such analyses ignore two channels that significantly influence the impacts of environmental policies. First, they disregard the budgetary impacts of environmental policies and the extent to which governments need to rely on other, distortionary taxes for revenue. Second, they ignore interactions between environmental policy initiatives and the functioning of markets outside of the market that is targeted by the environmental regulation. In particular, they disregard how the costs of environmental policies are influenced by pre-existing distortions in "other"

markets, including prior distortions caused by the tax system. The analyses described in this chapter indicate that these general-equilibrium considerations have first-order importance. Indeed, some of the most important impacts of environmental policies take place outside of the market that is targeted for regulation.

Sandmo (1975) was the first to consider general-equilibrium interactions in his important contribution analyzing optimal commodity taxation when one of the commodities involves an externality. More recent work shows that such interactions have economic implications that extend beyond those revealed in Sandmo's seminal analysis. The recent work emphasizes two fundamental ideas. First, environmental taxes and other forms of environmental regulations act as implicit taxes on factors of production because they raise the costs and prices of produced goods relative to the prices of factors, thereby lowering real factor returns. Second, these implicit taxes typically compound the distortions posed by pre-existing explicit factor taxes.

These two notions have profound implications for a number of issues in environmental regulation – for the costs of revenue-neutral environmental tax reforms, for the optimal environmental tax rate, and for the choice between environmental taxes and other instruments for environmental protection. They imply that, in many circumstances, the gross costs [1] of environmental tax reforms are higher in a second-best world than in a first-best setting. (This does not remove the efficiency rationale for environmental taxes: this chapter's analyses show that environmental taxes still can produce significant net efficiency gains, once environment-related benefits are taken into account. But prior distortionary taxes tend to imply higher gross costs than otherwise would be the case.) These two notions also imply that the optimal rate of tax on environmentally harmful activities is likely to be less than the rate endorsed by Pigou – that is, less than the marginal external damages. In addition, they imply that pollution taxes may have a significant potential advantage over (non-auctioned) pollution quotas because only the former raise revenue that can finance cuts in pre-existing distortionary taxes, thereby avoiding some of the efficiency costs that such prior taxes generate.

We explore these issues in detail in the rest of this chapter, which is organized as follows. Section 2 examines the optimal level of environmental taxes, both in a first-best setting and in a second-best setting where the government needs to impose distortionary taxes to generate revenues. Section 3 analyzes the impacts of environmental tax reforms, concentrating on revenue-neutral policies in which revenues from environmental taxes are used to finance cuts in ordinary, distortionary taxes. Here we consider the welfare implications of such policies, decomposing the impacts into environment-related benefits and non-environment-related costs. An issue of particular interest is the circumstances under which the "recycling" of revenues from environmental taxes through cuts in distortionary taxes can eliminate

[1] Gross costs are the costs before netting out the benefits associated with the policy-related improvement in environmental quality.

the non-environmental costs of such reforms. Whether green tax reforms can be introduced at zero cost has become a controversial issue in recent years. Section 4 compares environmental taxes with other policy instruments – emissions quotas, performance standards, and subsidies to abatement – in a second-best setting with pre-existing distortionary taxes. We first compare these instruments assuming that policymakers face no uncertainties as to firms' abatement costs or the benefits of environmental improvement. We then expand the analysis to consider how uncertainty on the part of regulators and associated monitoring and enforcement costs affect the choice among alternative policy instruments. Section 5 concentrates on the trade-offs between efficiency and distribution that arise in a second-best setting. Section 6 offers conclusions.

2. Optimal environmental taxation

This section employs simple general-equilibrium models to examine the optimal rate of tax on environmentally damaging activities. We first pose this issue as a planner's problem, and then consider how the government can produce the optimum in a decentralized market economy. We will consider these issues both in a first-best setting (devoid of pre-existing distortions) and in a more realistic setting involving other, distortionary taxes.

2.1. The basic model

Consider a representative household that derives utility $U = u(C, D, V, G, Q)$ from private goods – namely, a "clean" private good (C), a "dirty" private good (D), and leisure (V) – and from two public goods – non-environmental (i.e., produced) public goods (G) and the quality of the environment (Q). We apply the label "dirty" to those goods or services whose production or consumption directly contributes to deterioration of the environment. Let production be described by the constant-returns-to-scale production function $F(NL, X, R)$ for which the inputs are aggregate labor (the product of the number of households, N, and per capita labor supply, L), a "clean" intermediate good (X), and a "dirty" intermediate good (R). Gross output can be used to provide (non-environmental) public goods, to meet demands for clean or dirty intermediate inputs, or to meet household demands for clean or dirty consumption goods. Thus, the material-balance condition for the economy is

$$F(NL, X, R) = G + X + R + NC + ND. \tag{1}$$

We have normalized units so that the constant rates of transformation between the five produced commodities are unity.

Environmental quality, Q, deteriorates with the quantity used of dirty intermediate and dirty consumption goods:

$$Q = q(R, ND), \quad q_R, q_{ND} < 0. \tag{2}$$

Throughout, a subscript to a function will denote a partial derivative with respect to a given variable. Each household has one unit of time available that can be used for either work (L) or leisure (V):

$$V + L = 1. \tag{3}$$

2.2. *The first-best solution in a command economy*

The first-best outcome can be attained in a command economy. The social planner's objective is to maximize the utility of the representative household subject to conditions (1), (2), and (3). This constrained optimization problem yields the following first-order conditions:

$$u_C = u_V/F_{NL} = Nu_G = u_D + Nu_Q q_{ND}, \tag{4}$$

$$F_X = 1 = F_R + Nu_Q q_R/u_C. \tag{5}$$

Expression (4) indicates that u_V/u_C, the marginal rate of substitution between leisure and clean consumption, must equal the economy's marginal rate of transformation, which in this model is F_{NL}, the marginal product of labor. In addition, the marginal utility of C should equal the marginal social value of G and of D. The marginal social value of the (nonrival) public good G is the incremental utility, summed over the N households. The marginal social value of D is u_D plus a (negative) term correcting for the environmental damages associated with producing or consuming D (note $q_{ND} < 0$). The latter term is the sum of individual environmental damages over the households, reflecting the nonrival nature of environmental quality. Thus, the incremental social value of D involves both private-good and public-good elements, since greater provision and use of D affects utility not only through its private consumption but also by reducing Q, the "environmental public good".

Expression (5) governs the optimal use of intermediate inputs. At the optimum, the marginal product of the clean intermediate input X equals the marginal rate of transformation (i.e., unity). The marginal product of the dirty intermediate input R, in contrast, exceeds the marginal rate of transformation by an amount representing the marginal environmental damage associated with using this input.

2.3. *First-best outcome in a decentralized market economy: the "Pigouvian result"*

If the government has access to a sufficient set of policy instruments, it can achieve the first-best outcome in a decentralized market economy. To obtain this outcome, the

government must be able not only to impose taxes on goods but also to employ lump-sum taxes or subsidies. With these instruments at its disposal, the government's budget constraint amounts to

$$T + G = t_X X + t_R R + t_C NC + t_D ND + t_L wNL, \tag{6}$$

where w is the after-tax wage and t_i $(i = X, R, C, D, L)$ represents the tax rates on the transactions of i. The labor tax rate t_L is an ad-valorem tax on wages. T represents lump-sum transfers provided by the government to each household.

In a decentralized economy, households face the following budget constraint:

$$(1 + t_C)C + (1 + t_D)D = wL + T. \tag{7}$$

Private agents ignore environmental externalities when they implement their decentralized decisions. Accordingly, maximizing utility subject to the budget constraint (7) involves the following optimality conditions for the representative household:

$$u_C = u_V \frac{1 + t_C}{w} = u_D \frac{1 + t_C}{1 + t_D}. \tag{8}$$

Under perfect competition, firms maximize profits by equating the marginal product of each factor to its user cost:

$$F_{NL}(1, X/NL, R/NL) = w_p, \tag{9}$$

$$F_X(1, X/NL, R/NL) = 1 + t_X, \tag{10}$$

$$F_R(1, X/NL, R/NL) = 1 + t_R, \tag{11}$$

where w_p equals $w(1 + t_L)$ and represents the producer wage.

The government can establish the first-best outcome by levying taxes on the dirty intermediate and dirty consumption good. The first-best production condition (5) can be obtained if the government imposes a tax on the dirty intermediate good at a rate given by

$$t_R = t_R^P \equiv \frac{N u_Q(-q_R)}{u_C}. \tag{12}$$

The first-best optimal value of the tax rate t_R is equal to the social cost associated with the environmental harm from an increment in R. We define t_R^P as this marginal environmental harm. t_R^P is often referred to as the "Pigouvian" tax rate, after Pigou (1938), who articulated the idea that taxes could be used to generate an efficient outcome by "internalizing" environmental costs. Without government intervention, there is a wedge between marginal social and private cost. Decentralized firms consider only private cost (that is, the market prices of inputs), ignoring the environmental component of the social cost associated with the use of the dirty intermediate input.

The Pigouvian tax serves to eliminate the cost wedge, raising private cost to a level that corresponds to social cost. In this way, the Pigouvian tax internalizes the social cost from pollution.

Equation (12) can be interpreted as the condition for the optimal provision of the environmental public good. Note that this condition resembles the well-known "Samuelson condition" [see Samuelson (1954)] for the optimal provision of the (non-environmental) public good G. The Samuelson condition is implied by Equation (4), and can be written as

$$1 = \frac{Nu_G}{u_C}. \tag{13}$$

Expression (12) equates marginal social costs and (environmental) benefits related to a reduction in R, while Equation (13) equates marginal social costs and benefits related to an increase in G. In both (12) and (13) the right-hand side expresses the benefits in terms of the sum, over the N households, of the marginal rates of substitution between the public good involved and clean private goods. The left-hand side of Equation (13) represents the social cost of a one-unit increase in G. This is the marginal rate of transformation between private and produced public goods (i.e., unity). The left-hand side of Equation (12) stands for the social cost of improving environmental quality through a one-unit reduction in the use of R. This social cost corresponds to the loss of tax revenue associated with this one-unit reduction; hence this cost is simply the tax rate, t_R. Thus, the optimal value for t_R is the marginal social benefit from the environmental improvement stemming from a one-unit reduction in R.

The government can induce households to make efficient decisions by having the three tax rates t_C, t_D and t_L meet the following two conditions:

$$(1 + t_C)(1 + t_L) = 1, \tag{14}$$

$$(1 + t_D) = \left(1 - \frac{Nu_Q q_{ND}}{u_C}\right)(1 + t_C). \tag{15}$$

As in other optimal tax models with constant-returns-to-scale production functions, the government has one degree of freedom in setting these tax rates. This occurs because the household budget constraint (7) is unaffected if both expenditures (the left-hand side) and income (the right-hand side) are multiplied by the same factor. With one degree of freedom, we can choose to normalize the tax system by setting one of the tax rates to zero and solving for the other two[2]. In the rest of this chapter

[2] This degree of freedom implies that the first-best equilibrium can be achieved, in principle, even if administrative or political constraints prevent the government from introducing the tax t_D on dirty consumption. In this event, the first-best can be achieved through the combination of a subsidy to clean goods equal to $t_C = N(u_Q q_{ND}/u_C)/[1 - Nu_Q q_{ND}/u_C)]$ and a labor tax equal to $t_L = -N(u_Q q_{ND}/u_C$. Fullerton (1997) points out that this solution – a labor tax plus a subsidy on clean consumption – resembles a deposit-refund system.

we normalize the tax system by selecting the clean consumption good as the untaxed commodity (i.e. $t_C = 0$). Under our normalization, t_D isolates the differential taxation of the dirty consumption good. This is the taxation over and above the implicit taxation of all consumption from the labor tax, which is equivalent to a uniform tax on both consumption commodities. When one refers to pollution taxes, one typically has this tax differentiation in mind.

The government can produce the first-best outcome by refraining from taxing labor and by setting the tax on dirty consumption equal to the Pigouvian tax:

$$t_D = t_D^P \equiv \frac{Nu_Q(-q_{ND})}{u_C}. \tag{16}$$

The first-best solution also requires determining the optimal quantity of G and the optimal level of lump-sum transfers T. The Samuelson rule (13) determines the optimal quantity of G. Optimal lump-sum transfers are given residually from the government budget constraint (6).

2.4. When lump-sum taxes are not available: the second-best optimum

In practice, lump-sum taxes and subsidies typically are not available because of political and administrative constraints. Under such circumstances, the government's problem is to select values for its five fiscal instruments (t_L, t_D, t_X, t_R and G)[3] in order to optimize household utility subject to the government budget constraint and decentralized optimizing decisions by firms and households. The Lagrangian function is therefore:

$$NW((1+t_D), w, G, q(R, ND)) + \mu\,[t_L wNL + t_D ND + t_X X + t_R R - G]. \tag{17}$$

Here W represents indirect utility, and μ denotes the marginal disutility of raising one additional unit of public revenue. The optimization problem yields the following optimal tax rates on intermediate inputs [see Bovenberg and Goulder (1996)]:

$$t_X = 0, \tag{18}$$

$$t_R = \left[\frac{Nu_Q(-q_R)}{u_C}\right]\frac{1}{\eta}, \tag{19}$$

where η is defined as μ/λ, with λ representing the marginal utility of private income. Thus, η is the ratio of the shadow cost of raising government revenue to the shadow value of an incremental increase in private income. This is usually referred to as the marginal cost of public funds (MCPF)[4].

[3] Recall that we normalize the tax system by setting the tax rate on consumption equal to zero.

[4] In the presence of distortionary taxes, the MCPF depends on the choice of the untaxed good [see Boadway and Keen (1993)]. By selecting clean consumption as the untaxed good, we measure the MCPF in terms of clean consumption.

In the rest of this Subsection we assume that environmental quality is weakly separable from other goods in utility; that is, $U = u(P(C,D,V,G),Q)$. The more general case is explored in Subsection 2.5.1. With this utility function, the first-order conditions with respect to t_L, t_D and G yield the following expressions:

$$(\lambda - \mu) L + \mu \left[(t_D - t_D^Q) \frac{\partial D}{\partial w} + t_L w \frac{\partial L}{\partial w} \right] = 0, \tag{20}$$

$$(\lambda - \mu) D - \mu \left[(t_D - t_D^Q) \frac{\partial D}{\partial t_D} + t_L w \frac{\partial L}{\partial t_D} \right] = 0, \tag{21}$$

$$N \left(\frac{u_G}{u_C} \right) = \eta \left[1 - (t_D - t_D^Q) N \left(\frac{\partial D}{\partial G} \right) - t_L N w \left(\frac{\partial L}{\partial G} \right) \right], \tag{22}$$

where

$$t_D^Q \equiv \left[\frac{N u_Q(-q_{ND})}{u_C} \right] \frac{1}{\eta}. \tag{23}$$

We shall refer to these expressions in the discussion of optimal taxes below.

2.4.1. Optimal taxes on intermediate inputs

Expression (18) reveals that the clean intermediate input should not be subject to any tax. This is an application of the well-known optimality of production efficiency derived by Diamond and Mirrlees (1971). They demonstrated that, if production exhibits constant returns-to-scale, an optimal tax system should not distort production, that is, it should not directly alter the relative prices of intermediate inputs[5]. Intuitively, consumer taxes can yield the same effects on relative prices as a tax on intermediate inputs. Thus a tax on intermediate inputs does not provide any benefits relative to consumer taxes in terms of changes in relative consumer prices. At the same time, a tax on intermediate inputs introduces additional inefficiencies relative to optimal consumer taxes because it distorts relative input prices. Accordingly, consumer taxes dominate taxes on (clean) intermediate goods.

Expression (19) indicates that the tax on the dirty intermediate input, t_R, should be positive as long as households value environmental quality (i.e., $u_Q > 0$). The term in square brackets on the right-hand-side of Equation (19) corresponds to the Pigouvian tax [see Equation (12)]. The Pigouvian tax is optimal only if the marginal cost of public funds, η, equals unity. A unitary MCPF means that obtaining a dollar of public revenue involves, in general equilibrium, a one-dollar sacrifice of private income. In a second-best world without lump-sum taxation, the MCPF typically differs from one. Hence the

[5] Under decreasing returns to scale, production efficiency continues to be optimal so long as a 100 percent profit tax is available.

MCPF term in Equation (19) indicates how second-best considerations affect optimal environmental taxation. In particular, the higher the MCPF, the smaller the optimal environmental tax, *ceteris paribus.*

The optimal environmental tax is inversely related to the MCPF for the following reason. The government employs the tax system to simultaneously accomplish two goals: raising revenues and internalizing environmental externalities. Environmental taxes directly affect both objectives. If raising public revenues becomes more costly, as indicated by a higher MCPF, the balance between the revenue and environmental-quality objectives is best struck at a lower rate for the environmental tax. Specifically, the optimal pollution tax must balance the marginal social benefit from one unit of pollution reduction [the term in brackets on the right-hand side of Equation (19)] against the gross marginal social cost of a one-unit pollution reduction. The latter is the social cost associated with the reduction in pollution-tax revenue from a one-unit reduction in pollution. This, in turn, is equal to the MCPF times the pollution tax rate. Dividing these marginal benefits and costs by the MCPF gives Equation (19). Therefore, the higher the social cost of raising revenue, the higher the marginal social benefits from pollution abatement have to be to justify a given environmental tax. Thus, high estimates for the efficiency costs of existing taxes imply lower values for the optimal environmental tax rate.

To illustrate the analogy of environmental quality with other public consumption goods, we write Equation (22) for the case of a produced public good G that is weakly separable from private goods:

$$\frac{1}{\eta}\left(\frac{Nu_G}{u_C}\right) = 1. \tag{24}$$

The right-hand side of Equation (24) is the marginal rate of transformation between private and public goods. The MCPF drives a wedge between this rate of transformation and the sum of the marginal rates of substitution. A higher MCPF means that higher marginal benefits from public consumption are necessary to offset the higher efficiency cost of financing this public good. This is analogous to the effect of the MCPF on the required marginal benefits from the environmental public good [see Equation (19)].

2.4.2. Optimal taxes on consumer goods

To explore the optimal taxes on labor and dirty consumption (i.e., t_L and t_D), we first derive "Ramsey tax rules". These rules yield the least distortionary way of financing public spending if environmental externalities are absent (i.e., when $u_Q = 0$). We then turn to the more general policy problem in which taxes face the dual task of not only generating revenues to finance public spending but also internalizing environmental externalities.

2.4.2.1. Ramsey tax schemes. Without environmental externalities (i.e., with $t_D^Q = 0$), Equations (20) and (21) can be solved [see Bovenberg and van der Ploeg (1994b)] to yield the following:

$$\frac{t_D}{1+t_D} = \left(\frac{\varepsilon_{CL} - \varepsilon_{DL}}{\varepsilon_{CD} - \varepsilon_{DD}}\right) t_L, \tag{25}$$

where ε_{ik} stands for the compensated elasticity of demand for commodity i with respect to the price of commodity k. Optimal government policy thus involves both a tax on labor (which is equivalent to an equal tax on both consumption goods) and a tax or subsidy on the "dirty" consumption good[6]. The combination prescribed by Equation (25) is equivalent to a set of taxes on the two consumption goods, with a different tax rate applying to the dirty consumption good. In the absence of externalities, the tax on the dirty consumption good is a Ramsey tax; it is motivated purely by non-environmental considerations. The sign of this tax depends on the cross-elasticities with leisure. In particular, the tax rate is positive if $\varepsilon_{CL} > \varepsilon_{DL}$, that is, if the clean consumption good is a better substitute for leisure than the dirty consumption good is. In that case, the dirty consumption good is a relative complement to leisure. Thus it is optimal for the government to levy (via the labor tax) a uniform tax on clean and dirty consumption goods, and to supplement this with a tax on the good that is most complementary to leisure[7].

2.4.2.2. Integrating Ramsey and Pigou. We now turn to the case with environmental externalities. In the presence of externalities, t_D^Q is non-zero, and thus in Equations (20–22) the term $(t_D - t_D^Q)$ replaces what was simply t_D when externalities were absent. Now the tax t_D has both a Ramsey (or distortionary) component and an environmental (or non-distortionary) component. The term $(t_D - t_D^Q)$ is the Ramsey component. It follows that the optimal tax rate is the sum of the Ramsey and externality-correcting terms[8]:

$$\frac{t_D}{1+t_D} = \left(\frac{\varepsilon_{CL} - \varepsilon_{DL}}{\varepsilon_{CD} - \varepsilon_{DD}}\right) t_L + \frac{t_D^Q}{1+t_D}. \tag{26}$$

The first part of the optimal pollution tax on consumption [i.e., the first term on the right-hand side of Equation (26)] is the Ramsey component of the tax on

[6] We retain the label "dirty" despite the assumption here that there are no environmental externalities.

[7] See also Corlett and Hague (1953), and Diamond and Mirrlees (1971, p. 263) for an expression analogous to Equation (25) and a related discussion.

[8] See Sandmo (1975). Ng (1980) explores the sign of the optimal pollution tax. He finds that, in the presence of environmental externalities (i.e. $u_Q > 0$), the pollution tax is typically positive. However, if the revenue requirement is small and falls short of the revenues from the Pigouvian tax, the optimal pollution tax may actually be negative. In this counterintuitive case, a lower consumption wage must be very effective in reducing dirty consumption, compared to a higher consumption price for dirty consumption. Hence, the combination of a wage tax and a subsidy on dirty consumption reduces pollution.

polluting consumption. Together with the optimal labor tax, the optimal level of the Ramsey component is determined on the basis of the familiar Ramsey formulas for raising revenues with the lowest costs to private incomes [see Equations (20) and (21)]. This component measures the social contribution (in terms of government revenues) of additional demand for the dirty consumption good as the difference between a positive and a negative contribution. On the one hand, consumption of the dirty consumption good boosts the tax base and thus facilitates the financing of ordinary public goods. On the other hand, it damages the environment, thereby reducing the supply of the environmental public good.

The second part of the optimal pollution tax is t_D^Q [Expression (23)]. This part corrects for the environmental externality. The expression for t_D^Q looks very similar to the expression for the optimal tax t_R on the dirty intermediate input [see Equation (19)]. It is the Pigouvian tax divided by the MCPF. Using Equation (20), we can write the MCPF as

$$\eta = \left(\frac{1}{1 - (t_D - t_D^Q)(D/wL)\, \varepsilon_{DL}^U - t_L \varepsilon_{LL}^U} \right), \tag{27}$$

where ε_{ik}^U stands for the uncompensated elasticity of demand for commodity i with respect to the price of commodity k. The MCPF exceeds unity if financing additional public spending erodes the base of existing Ramsey (or distortionary) taxes[9].

2.4.2.3. A special case. To generate further insights, we derive results for the particular case where the utility function is homothetic, and clean and dirty consumption are weakly separable from leisure. This implies that the compensated elasticities ε_{CL} and ε_{DL} are identical. In this case the Ramsey-tax term in Equation (26) is zero and the pollution tax reduces to the externality-correction term (i.e., $t_D = t_D^Q$). In this special case, the MCPF can be written as [from Equation (27) with $t_D = t_D^Q$]:

$$\eta = \left[1 - t_L \varepsilon_{LL}^U\right]^{-1}. \tag{28}$$

The MCPF thus exceeds unity if, first, the uncompensated wage elasticity of labor supply, ε_{LL}^U, is positive and, second, the distortionary tax on labor, t_L, is positive. The latter condition holds when Pigouvian taxes are not sufficient to finance the optimal level of public consumption. These results are consistent with the literature

[9] If taxed commodities are inferior, the MCPF may actually fall short of unity. The reason is that the negative income effect associated with a higher tax level may raise the consumption of taxed commodities. The MCPF is smaller than unity if, in the terminology of Atkinson and Stern (1974), the "revenue effect" of a tax increase boosts the demand for taxed commodities and is large enough to more than offset the "distortionary effect" of tax increases. See also Section 5.1 in Chapter 21 by Auerbach and Hines in this volume, and Ballard and Fullerton (1992).

on the MCPF surveyed in Ballard and Fullerton (1992). In the case where utility from public consumption is separable from consumer's choice on leisure and consumption, this literature finds that distortionary labor taxes raise the marginal costs of public spending above unity if the uncompensated wage elasticity of labor supply is positive. The same condition on this uncompensated elasticity determines whether distortionary labor taxes raise the marginal cost of the environmental public good above its social benefit [see Equations (19), (23) and (28)].

2.4.2.4. Optimal level of public consumption. The adjusted Samuelson rule (22) indicates that there are two reasons why the marginal rate of transformation between private and public goods differs from the corresponding sum of the marginal rates of substitution. The first is that the marginal cost of public funds (η) may differ from unity. As indicated in Expression (27), raising additional government revenue may cause an erosion of the base of pre-existing distortionary taxes, thereby imposing costs over and above the revenue collected from the new tax. In such circumstances the marginal cost of public funds exceeds unity.

The second reason for the divergence between the marginal rates of transformation and substitution is that, if public goods are complementary to taxed commodities ($t_D \partial D/\partial G > 0$ or $t_L \partial L/\partial G > 0$), raising public spending alleviates the excess burden of distortionary taxation by boosting the consumption of taxed commodities. For example, the construction of public highways between suburbs and cities may induce some agents to work more and thus pay more labor tax. Public libraries, in contrast, may encourage private agents to enjoy more leisure, thereby eroding the base of the labor tax. For this reason the social cost of funds devoted to libraries can exceed the cost of the same amount of funds allocated to highways.

2.5. Some complications to the second-best problem

2.5.1. The environment as a non-separable consumption good

Thus far, we have assumed that environmental quality is separable in utility from consumption and leisure. If this is not the case, environmental quality directly affects private decisions and the optimal non-distortionary component of the tax on dirty consumption is given by [see Bovenberg and van der Ploeg (1994b)]:

$$t_D^Q = \frac{N(-q_{ND})}{u_C \eta} \left(\frac{u_Q + \mu \left[t_D \frac{\partial D}{\partial Q} + w t_L \frac{\partial L}{\partial Q} \right]}{1 - N \frac{\partial D}{\partial Q} q_{ND}} \right). \tag{29}$$

Equation (23) showed that when the MCPF differs from unity, the optimal environmental tax t_D^Q differs from the sum of the marginal rates of substitution. When the environment is non-separable in utility, an additional factor contributes to a difference between t_D^Q and the sum of marginal rates of substitution. In particular, if the environment is a

gross complement to leisure (i.e., if $\partial L/\partial Q < 0$), then improvements in environmental quality come at a higher cost because the environmental tax leads to a greater reduction in the labor tax base. [See the numerator of the far-right term in Equation (29)]. In this case, the social value of environmental protection is reduced and the optimal environmental tax falls [10].

The denominator of the term in large brackets in Equation (29) accounts for environmental quality's "feedback effect" on the demand for dirty goods. In particular, if an improvement in environmental quality raises the demand for dirty goods (i.e. $\partial D/\partial Q > 0$), the net benefit from increased environmental quality is reduced. Traffic congestion illustrates this case. Less traffic congestion encourages more traffic. Accordingly, while higher taxes on gasoline reduce congestion, the overall impact of these taxes on congestion is mitigated by the feedback of reduced congestion on traffic [11].

2.5.2. The environment as a public input to production

The foregoing analysis treats environmental quality as a public consumption good. Amenities like clean air, relative quiet, and greater visibility fall into this category. However, environmental quality also functions as a public input into production. For example, since certain types of agricultural production benefit from a cooler climate, slowing down global warming can avoid some losses of agricultural productivity. Furthermore, reduced air pollution is likely to improve health and thereby boost labor productivity. To model the impact of the environment on production, we specify production as follows:

$$Y = a(Q)F(NL, X, R); \qquad a'(Q) > 0, \quad a''(Q) \leqslant 0. \tag{30}$$

With this formulation, the expression for the optimal tax rate on dirty inputs becomes [see Bovenberg and van der Ploeg (1994a)]:

$$t_R = \left[\frac{Nu_Q(-q_R)}{u_C}\right]\frac{1}{\eta} + (-q_R)\,a'(Q)\,F. \tag{31}$$

The first term on the right-hand side matches the one that applies when the environment is only a consumption good [see Expression (19)]. This term represents

[10] These findings parallel results obtained in the literature on the optimal supply of ordinary (i.e., non-environmental) public goods in the presence of distortionary taxation. In that literature, the way a particular public good enters utility affects the marginal costs of financing such a public good. See Wildasin (1984) as well as Subsection 2.4.2 on the optimal level of public consumption.

[11] Cornes (1980) and Sandmo (1980) show that for the aggregate demand function to be stable, the feedback effect cannot be too large in absolute value. This stability condition ensures that the denominator of the last term on the right-hand side of Equation (29) is positive.

the consumption externality. The second term on the right-hand side represents the adverse effect of pollution on productivity. In contrast with the first term, the second term does not involve the marginal cost of public funds[12].

2.5.3. *Environmental damages from accumulated pollution stocks*

Thus far we have associated environmental quality with the current level of output of dirty goods. For certain types of pollution, where the flow of pollution does not contribute to a durable stock, environmental quality can be viewed as directly connected to the pollution flow. Noise pollution provides a pertinent example. But in most circumstances, environmental quality or damage is more closely connected to the stock of pollution, and in such cases the relationship between pollution emissions and environmental quality is inherently dynamic. These dynamic connections imply a more complex formulation of the optimal environmental tax rate, although this formulation still echoes the principles that apply when a simpler pollution–quality relationship is assumed.

Here we consider the optimal environmental tax rate in a model that relates environmental quality (or damages) to pollution concentrations. Our example is the climate-related economic damage associated with atmospheric accumulation of carbon dioxide (CO_2). The problem at hand is to obtain the optimal profile of taxes on CO_2 – carbon taxes – to maximize environmental gains net of abatement costs induced by the tax. The first analytical studies of this problem appear to be those by Nordhaus (1980, 1982)[13]. The problem can be viewed as maximizing the discounted stream of utility from consumption:

$$\max_{\{c(t)\}} U = \int_0^\infty \exp(-rt)\, u(c(t))\, \mathrm{d}t, \tag{32}$$

where $c(t)$ denotes consumption at time t and r is the utility discount rate. (c should be distinguished from the clean consumption good C, which appeared earlier.) At each

[12] Environmental regulations discourage labor supply by implicitly taxing labor – that is, raising the costs of consumption goods relative to leisure. At the same time, when the environment is a productive input such regulations promote greater labor supply by enhancing labor productivity (i.e., avoiding damages to production). As shown by Williams (1997), these two effects cancel out at the optimum, which implies that MCPF need not be considered in determining the contribution of the production-side effect to the optimal tax rate. See also Eskeland (2000). This result reaffirms Diamond and Mirrlees' (1971) finding that production efficiency is optimal.

[13] For other analytical treatments, see Sinclair (1994), Ulph and Ulph (1994), Peck and Wan (1996) and Goulder and Mathai (2000). Nordhaus (1994), Peck and Teisberg (1994), Manne and Richels (1992), Farzin and Tahvonen (1996) and several other authors have employed simulation models to solve numerically for optimal carbon tax profiles.

point in time, consumption depends on $e(t)$, current CO_2 emissions, and on $S(t)$, the current atmospheric concentration of CO_2:

$$c(t) = f[e(t)] - h[S(t)]. \tag{33}$$

The function f indicates that abating emissions involves economic costs that translate (other things equal) into a loss of consumption; the function h indicates that increases in the stock of CO_2 affect climate patterns and thereby reduce consumption. The evolution of the CO_2 stock is given by

$$\dot{S}(t) = \alpha e(t) - \delta S(t), \tag{34}$$

where α and δ are parameters. The solution to this problem [see Nordhaus (1982)] is:

$$-\Omega(t) = f'[e(t)] = \frac{\int_t^{\infty} \alpha e^{-(r+\delta)} u'[c(s)]\, h'[S(s)]\, ds}{v(t)}, \tag{35}$$

where $\Omega(t)$ (a negative number) is the shadow value of the stock of CO_2 at time t, and $v(t)$ is the shadow value of consumption at time t. Expression (35) indicates that, at the optimum, the CO_2 shadow price is equal to both the marginal cost of reducing emissions (the expression to the right of the first equality sign) and the discounted cost of the change in atmospheric concentration stemming from a marginal increase in emissions (the expression to the right of the second equality sign). The latter is equivalent to the marginal benefit from incremental emissions reductions. Thus, at the optimum, marginal costs and benefits of emissions reductions are equated. If the government sets the carbon tax equal to marginal benefits from incremental emissions reductions (or marginal damages from incremental emissions), it will satisfy the second condition in Equation (35). If producers are competitive, they will equate marginal costs of abatement to the carbon tax, thereby satisfying the first condition in Equation (35). Thus, the optimal carbon tax profile involves setting carbon taxes equal to the negative of $\Omega(t)$ at all periods of time.

The negative of the shadow price of carbon concentrations (or optimal carbon tax) evolves according to

$$-\dot{\Omega}(t) = (r + \delta)[-\Omega(t)] - h'(S(t)). \tag{36}$$

Other things equal, a higher discount rate r necessitates a faster increase in the optimal carbon tax. Higher discount rates reduce the relative price of future abatement costs relative to current abatement costs. Hence more future abatement becomes justified, and the carbon tax must rise more quickly to encourage relatively more abatement (less emissions) in the future. A higher value for the "removal" rate δ implies a faster increase in the carbon tax. Higher values for δ mean that carbon decays more quickly in the atmosphere. If δ is positive, a one-unit increase in emissions in period t

accompanied by a one-unit reduction in emissions in some future period $t + s$ would imply a greater overall amount of dispersion and thus a lower carbon concentration S in all periods after $t + s$. Hence there is a value to postponing emissions reductions. This value is larger, the higher is δ. Hence a larger value of δ justifies greater relative abatement in the future and thus a more steeply rising carbon tax.

The right-hand term in Equation (36) indicates a relationship between the slope of the damage function h and the growth of the optimal carbon tax. The more $h(S)$ is increasing in S, the greater the value to postponing emissions of CO_2, since postponed emissions delay the augmentation of the stock and thus imply a smaller present discounted value of damages. Thus larger values for h' imply a more slowly rising carbon tax profile, *ceteris paribus.*

This analysis of optimal tax rates disregarded second-best considerations. Such considerations do not fundamentally alter the results. The essential difference is that in a second-best setting the optimal carbon tax should equal marginal damages divided by the marginal cost of public funds, rather than simply the marginal damages. The optimal carbon tax still exhibits a profile similar to that given by Equation (36), assuming that the marginal cost of public funds does not change much through time. Second-best considerations imply a lower path for the optimal carbon tax, but the shape of the path need not differ much from that suggested by Equation (36).

2.6. Empirical issues and assessments

2.6.1. Marginal environmental damages

The foregoing analysis indicates that optimal pollution tax rates should reflect two main elements: the marginal environmental damages from pollution, and the MCPF. The model presented above did not reveal explicitly the complex connections between the use of certain inputs or products associated with pollution ("dirty" intermediate inputs or consumption goods) and the ultimate damages to the environment. Usually, several connections are involved: (1) from the use of an input or product to emissions of given pollutants, (2) from emissions of pollutants to concentrations, and (3) from concentrations to environmental damages.

These connections suggest that it may be more effective, other things equal, to impose taxes on emissions of specific pollutants rather than on given inputs or products associated with pollution, since emissions are more closely linked to the ultimate environmental damages. However, it may be more costly to monitor emissions than the use of inputs or products. These considerations are discussed in more detail in Section 4[14].

[14] In some cases, emissions are strictly proportional to the use of a given input, and in this case there is little sacrifice involved in treating the use of the input as a proxy for emissions. The clearest example of this is the relationship between the use of inputs of fossil fuels and emissions of carbon dioxide. For

The second and third connections above imply that the marginal damages may not be a simple function of emissions. If the concentration–damage relationship is nonlinear, the impact of an additional unit of emissions of a given pollutant depends on the concentration of the pollutant. Moreover, in some cases pollutants interact, so that environmental damages depend on the mix of pollutants rather than individual concentrations. For example, the extent to which atmospheric concentrations of nitric oxides or volatile organic compounds contribute to respiratory problems depends on the mix of these pollutants (specifically, the amount of ground-level ozone produced by the mix) rather than on the individual concentrations. Finally, the relationship between emissions and concentrations (connection 2) can depend on geographical conditions and meteorological factors (prevailing winds, etc.).

All of this indicates a great deal of complexity and heterogeneity in the relationship between emissions of given pollutants and the marginal environmental damages. Any estimates of *average* damages per ton of pollutant therefore will mask a great deal of spatial and temporal variation. With this important caveat in mind, we present estimates in Table 1 of average marginal damages for four of the six "criteria" air pollutants subject to Federal regulation in the USA under the 1990 Clean Air Act Amendments.

2.6.2. Estimates of the MCPF

The other key element in determining optimal tax rates is the MCPF. Optimal tax rates depart from the Pigouvian rates (that is, from the marginal environmental damages) to the extent that the MCPF differs from unity. A condition for an optimal tax system is equality in the MCPF for all taxes that generate government revenue. However, empirical estimates of the MCPF are obtained from tax systems that are generally suboptimal from an efficiency point of view. As a result, the estimates of the MCPF span a wide range, not only because studies employ different data and methodologies, but also because they reveal significant variations in the MCPF depending on the particular source of revenue involved. Table 2 provides a sampling of MCPF estimates.

These results suggest that for the USA environmental tax rates should be about 20–50 percent below the marginal environmental damages. The Hansson–Stuart results indicate that Swedish rates should be a smaller fraction of marginal damages.

2.6.3. Constrained-optimal policies: the case of the carbon tax

Bovenberg and Goulder (1996) consider these second-best issues in assessing optimal carbon taxes. Using a numerically solved intertemporal general equilibrium model of the USA, they find that the optimal carbon tax rate tends to be approximately

virtually all uses of fossil fuels or refined fossil-fuel products, the release of CO_2 into the atmosphere is strictly proportional to the carbon content of the fossil fuel or the refined product. Thus, a carbon tax is an excellent proxy for a tax on CO_2 emissions. Such proportionality, however, tends to be the exception rather than the rule.

Table 1
Marginal damages from pollution in the USA[a]

(1) Pollutant	(2) Major associated damages	(3) Region for which damage is calculated	(4) Marginal damage (1995 dollars per ton)	(5) 1996 Emissions (millions of short tons)	(6) Damage as percent of GDP in 1996[b]
Volatile organic compounds (VOCs)	health effects, agricultural losses, worsened visibility, materials damage	Northeastern USA	427–2828	19.09	0.107–0.707
Particulate matter	increased mortality, increased morbidity, soiling	USA	505–12 819	3.29[c]	0.022–0.552
Sulfur oxides	health effects, agricultural losses, worsened visibility, materials damage, soiling	USA	375–2087	19.11	0.094–0.522
Nitrogen oxides	worsened visibility, acute health effects, agricultural losses, materials damage	USA	10–122	23.39	0.003–0.037

[a] Sources: Climate and Policy Assessment Division, Office of Policy, Planning & Evaluation, US Environmental Protection Agency; National Air Quality and Emissions Trends Report (EPA Document Number 454/R-97-013), US Environmental Protection Agency (1996).
[b] Assuming constant marginal damages; figures in this column apply numbers from column (5) to the ranges in column (4), and express this product as a percentage of 1996 GDP ($7636 billion).
[c] Particulate matter of size 10 μm or less. Excludes "miscellaneous and natural" emissions.

Table 2
MCPF estimates

Study	Taxes considered	Estimate (MCPF per dollar of revenue)
Browning (1987)	US taxes	1.32–1.47
Hansson and Stuart (1985)	Swedish taxes	1.69
Ballard, Shoven and Whalley (1985)	US taxes	1.17–1.56
Bovenberg and Goulder (1996)	US taxes	1.11–1.41

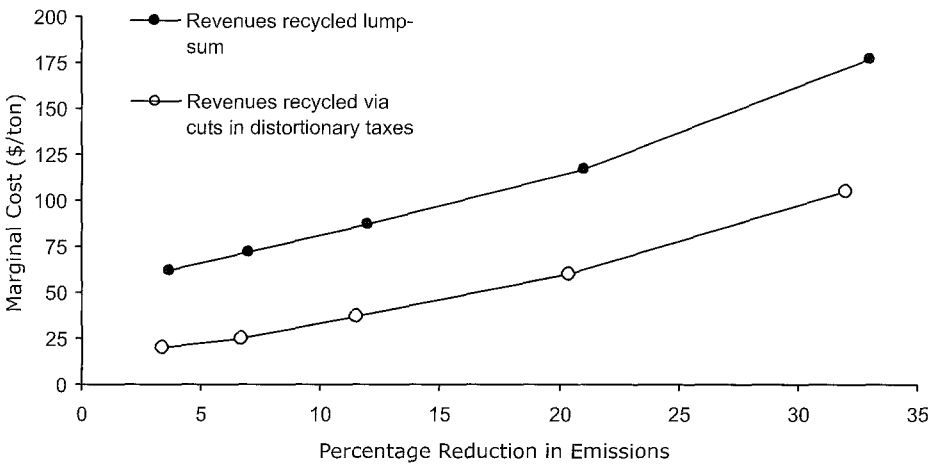


Fig. 1. General equilibrium marginal costs of carbon dioxide emissions reductions. Note: these are gross costs: they do not net out the benefits from avoided environmental damage.

20 percent below the marginal environmental damages (the central estimate of the MCPF is approximately 1.25).

They also consider a "constrained-optimal" carbon tax policy: a case where revenues from the tax are returned in lump-sum fashion rather than in the form of cuts in existing distortionary taxes (like the labor tax in the model above). In this case, the government forgoes an opportunity to avoid some of the distortionary costs imposed by existing taxes. This implies a higher schedule for the marginal costs of abatement relative to the case where revenues are returned through cuts in marginal rates, as shown in Figure 1.

The figure provides schedules for the general-equilibrium gross marginal cost of abatement under different assumptions about the use of the tax revenues. Here marginal cost is a gross concept in that it abstracts from the benefits associated with environmental improvement. The intercept of the marginal cost function is strictly positive for the case where revenues are returned lump-sum, whereas it is close to

zero for the case where revenues are returned through cuts in distortionary taxes [15]. In the scenario involving lump-sum recycling of the revenues, the positive intercept of this marginal cost function represents a threshold value for marginal environmental benefits from abatement: if marginal benefits are below this threshold, then any emissions abatement (or any positive carbon tax) is efficiency-reducing. Bovenberg and Goulder's central estimate for the intercept in the constrained-optimal case is about $50 per ton, which is higher than most estimates of marginal benefits from reducing carbon-dioxide emissions. Thus, failing to use revenues optimally can preclude efficiency gains from carbon taxation, and in this case a positive carbon tax is no longer efficiency-improving [16]. The use of revenues is also an important issue in the evaluation of environmental tax reforms, the subject of the next section.

3. Environmentally motivated tax reforms

This section considers reforms involving environmentally motivated taxes. Although proponents of policy reforms regard such changes as yielding improvements in economic outcomes, in contrast with the previous section the rates of environmental and other taxes examined here need not be optimal or even constrained-optimal.

Recently, environmental tax reforms have received increasing attention. One reason for this appears to be an increasing concern about environmental quality. A second and related reason is a growing recognition of options for substituting environmental taxes for other taxes as sources of revenue [17]. This latter issue is at the heart of most discussions of "green tax reform." Currently, revenues from environmental taxes represent on average about two percent of GDP and six percent of aggregate tax revenues in OECD countries (see Figure 2). Petroleum, diesel-fuel and motor-vehicle taxes account for most of these revenues (see Table 3). Proponents of green tax reform argue that society would benefit from increased reliance on environment-related taxes.

In examining environmental tax reforms, it will be useful to divide the welfare impacts into gross benefits and gross costs. The gross benefits are the gross welfare gains

[15] It is slightly positive because, according to the assumptions of the model, the US tax system is suboptimal on non-environmental dimensions. In particular, capital is overtaxed relative to labor. The intercept is zero under a counterfactual benchmark where labor, capital, and other non-environmental taxes are set optimally.

[16] The optimal carbon tax in this case is negative, assuming that the revenue cost of the negative tax is financed through lump-sum taxes. Just as a carbon tax implicitly raises factor taxes, a subsidy to carbon implicitly reduces such taxes. The implicit reduction in factor taxation yields a non-environmental efficiency improvement that more than offsets the efficiency loss associated with increased environmental damage. Of course, this policy is fairly unrealistic. If lump-sum taxes could finance a carbon subsidy, they might as well finance other aspects of government spending, making distortionary taxes unnecessary.

[17] Terkla (1984), Lee and Misiolek (1986), Ballard and Medema (1993) and Oates (1993) were among the first to analyze the potential efficiency benefits from using pollution-tax revenues to finance cuts in other, distortionary taxes. See Poterba (1993), Goulder (1994) and Oates (1995) for related discussions.

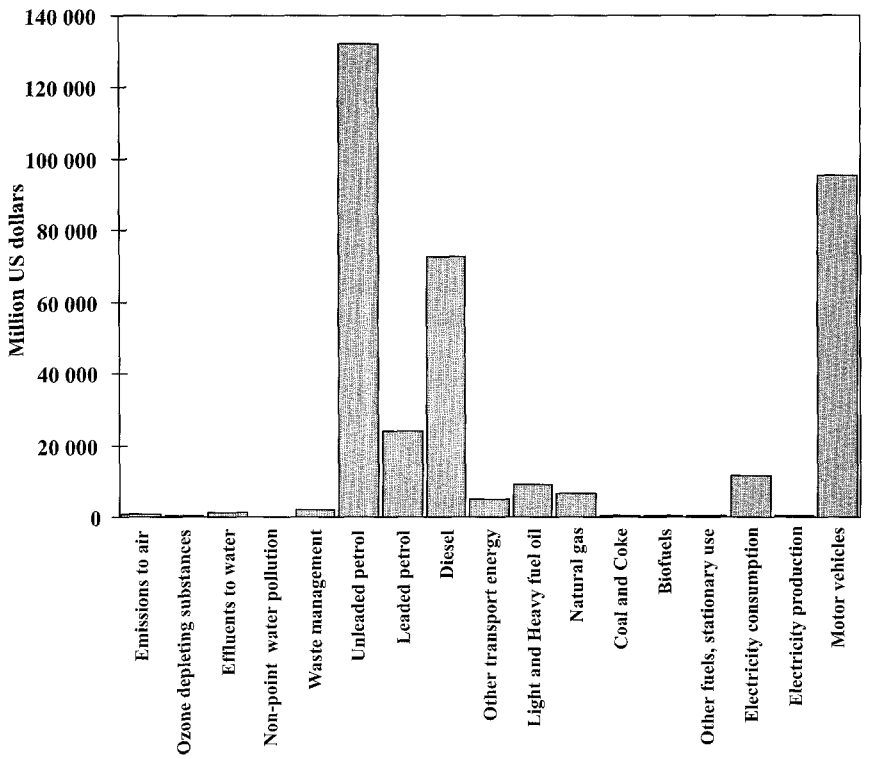


Fig. 2. Environment-related tax revenue in 21 OECD member countries in 1995.

associated with environmental improvement (or the avoidance of environmental deterioration). The gross costs are the gross welfare losses (ignoring environmental effects) associated with the economic sacrifices necessary to achieve reductions in pollution. A tax reform is efficiency-improving if it produces positive net benefits, that is, if gross benefits exceed gross costs.

Some analysts have suggested that replacing distortionary taxes by environmental taxes involves zero or negative gross costs. If this is so, such policies yield a "double dividend" by not only improving the environment but also reducing the non-environmental costs of the tax system. The interest in the second dividend reflects the political attractiveness of "no-regrets" policies: if the second dividend materializes, then environmental improvements can be produced with no cost to the economy. The interest also reflects the desires of policy analysts to justify policy reforms despite the significant uncertainties about the size of the first, environmental dividend. In the presence of the second dividend, the burden of proof facing policy makers is much reduced: to justify the environmental tax on benefit–cost grounds, it suffices to know that environmental benefits are non-negative. In this section we analyze the economic forces that determine the prospects for a double dividend. We show that

Table 3
Contributions of environment-related taxes to overall tax revenues for OECD countries in 1997[a]

Country	Tax revenue (millions of US dollars)		GDP (billions of US dollars)	Environment-related tax revenue as percent of	
	Environment-related	Total		Total tax revenue	GDP
Austria	4865	91297	206.7	5.33	2.35
Belgium	5715	111411	243.6	5.13	2.35
Canada	13242	236225	640.0	5.61	2.07
Czech Republic	1501	20460	53.0	7.33	2.83
Denmark	7780	84233	168.4	9.24	4.62
Finland	3963	56526	122.5	7.01	3.23
France	30156	635746	1406.0	4.74	2.14
Germany	46382	782305	2114.5	5.93	2.19
Greece	4746	40504	120.0	11.72	3.95
Hungary	1292	17868	45.8	7.23	2.82
Iceland		2377			
Ireland	2381	25772	78.5	9.24	3.03
Italy	37790	515237	1159.5	7.33	3.26
Japan	71388	1202355	4195.3	5.94	1.70
Korea	13333	101880	476.9	13.09	2.80
Luxembourg	504	7303	17.5	6.89	2.88
Mexico		67763			
Netherlands	13668	158109	376.7	8.64	3.63
New Zealand	1108	23553	64.9	4.70	1.71
Norway	5570	65676	155.0	8.48	3.59
Poland	2350	55936	143.2	4.20	1.64
Portugal	3670	34919	104.3	10.51	3.52
Spain	11964	188355	558.6	6.35	2.14
Sweden	7276	122252	237.5	5.95	3.06
Switzerland	5020	86729	256.3	5.79	1.96
Turkey	5846	53007	190.2	11.03	3.07
United Kingdom	38247	464383	1315.7	8.24	2.91
United States	77333	2299136	8121.0	3.36	0.95
Total	417090	7551318	22571.6	5.52	1.85

[a] Source: OECD.

although the double dividend is possible, it is unlikely to arise except under fairly unusual circumstances. More generally, this section examines how environmental taxes interact with other, distortionary taxes, and the implications of these interactions for the efficiency impacts of environmental reforms.

3.1. Gross costs and environment-related benefits of revenue-neutral reforms

We can evaluate the double dividend using the model from Section 2. Our approach will be to determine the welfare impacts of a revenue-neutral environmental tax reform, and to divide these impacts into environmental and non-environmental components. Consider the welfare effects of a revenue-neutral change in the tax mix (i.e., a change in taxes such that $dG = 0$). Taking the total differential of utility, we obtain

$$dU = u_C\,dC + u_D\,dD - u_V\,dL + u_Q q_R\,dR + N u_Q q_{ND}\,dD. \tag{37}$$

Substituting the first-order conditions for household optimization [from Equation (8)] into Equation (37), we can write

$$\frac{dU}{u_C} = dC + (1+t_D)\,dD - w\,dL + \frac{u_Q q_R\,dR}{u_C} + \frac{N u_Q q_{ND}\,dD}{u_C}. \tag{38}$$

Taking the total differential of goods–market equilibrium (1) and substituting the first-order conditions for profit maximization [i.e., Equations (9–11)], we find

$$N w_p\,dL + t_X\,dX + t_R\,dR = N\,dC + N\,dD. \tag{39}$$

Using Equation (39) to eliminate dC from Equation (38), we arrive at

$$\frac{dU}{u_C} = w t_L\,dL + \left[t_D - \frac{N u_Q(-q_{ND})}{u_C}\right] dD + \left[t_R - \frac{N u_Q(-q_R)}{u_C}\right] \frac{dR}{N} + t_X \frac{dX}{N}. \tag{40}$$

Equation (40) shows the welfare impacts associated with the changes in labor supply, input demands, and consumption. The first term on the right-hand side of Equation (40) stands for the distortionary effect in the labor market, which is regulated by the pre-existing tax on labor income. The next two terms correspond to the effects on the environmental margin. The welfare impact of a marginal increase in the demand for dirty goods amounts to the difference between a tax term, which measures the social benefits of additional tax revenue due to a wider revenue base, and the marginal social damage from pollution. When t_D and t_R are set at the Pigouvian tax rates [see Equations (16) and (12), respectively], each of the terms in square brackets is zero: beneficial environmental effects associated with less pollution exactly offset the adverse welfare effects due to an erosion of the tax base.

We can diagnose the welfare effects of tax changes by rearranging Expression (40):

$$\frac{\mathrm{d}U}{u_C} = -\frac{u_Q}{u_C}\left[N(-q_{ND})\,\mathrm{d}D + N(-q_R)\frac{\mathrm{d}R}{N}\right] + \left[wt_L\,\mathrm{d}L + t_D\,\mathrm{d}D + t_R\frac{\mathrm{d}R}{N} + t_X\frac{\mathrm{d}X}{N}\right]. \tag{41}$$

The product of $-u_Q/u_C$ and the first bracketed element on the right-hand side of Equation (41) represents the welfare effect of changes in environmental quality. The first "dividend" from environmental tax reform arises if this product is positive.

The other bracketed element on the right-hand side of Equation (41) stands for the welfare effect from changes in the tax base. This element is the *tax-base effect.* Each term contributing to this effect is the change in a tax base times the tax rate corresponding to that tax base[18]. This effect can be expressed as $\mathrm{d}Y^D$, the change in real private (after-tax) income enjoyed by households:

$$\mathrm{d}Y^D \equiv L\,\mathrm{d}w - D\,\mathrm{d}t_D = wt_L\,\mathrm{d}L + t_D\,\mathrm{d}D + t_R\frac{\mathrm{d}R}{N} + t_X\frac{\mathrm{d}X}{N}. \tag{42}$$

The tax-base effect represents the gross cost (i.e., the cost before netting out the environmental benefits) of the tax-induced changes in the allocation of resources. If this gross cost is negative, the environmental reform offers a second "dividend" in the form of a less costly tax system on non-environmental grounds. We now investigate the sign of the tax-base effect.

3.2. *Employment and welfare impacts of revenue-neutral reforms*

Consider in particular a reform in which the government introduces pollution taxes on household consumption or intermediate inputs and uses the revenues to finance cuts in the labor tax rate. We assume that the tax rate on the clean intermediate input is zero (i.e., $t_X = 0$)[19]. Utility is given by $u = u((M(J(C,D),V),G,Q)$. Hence, private goods are weakly separable from the public goods G and Q, so that environmental quality and public consumption do not directly affect private demand. The sub-utility function J aggregating clean and dirty consumption into a composite consumption good is homothetic. This specification of utility implies that, in the absence of environmental externalities, a uniform tax on clean and dirty consumption would be optimal.

Equation (42) implies that the non-environment-related welfare impact or gross cost of the reform depends on the reform's impact on labor supply. Because of this

[18] This formula for the first-order welfare change in the absence of externalities is standard in the tax-reform literature. Aronsson (1999) derives analogous expressions in a dynamic framework that considers the time-path of various endogenous variables in general equilibrium.

[19] As discussed in Section 2, there is no efficiency rationale for a tax on the clean intermediate input.

connection we first focus on the reform's impact on the base of the labor tax, that is, on employment[20].

To determine the general-equilibrium employment effects of the pollution tax t_D, we first derive $(u_D/u_C) = (1 + t_D)$ from household optimization. Taking the total differential, we find:

$$\tilde{C} - \tilde{D} = \sigma_J \tilde{t}_D, \tag{43}$$

where $\tilde{t}_D \equiv \mathrm{d}t_D/(1 + t_D)$. For other variables, a tilde ($\tilde{}$) stands for a relative change. σ_J represents the substitution elasticity between clean and dirty consumption in the sub-utility function $J(C,D)$. In deriving Equation (43), we have used the assumption that, in private utility, leisure is weakly separable from the produced commodities C and D. Under these separability assumptions and homotheticity of the sub-utility function $J(C,D)$, the first-order conditions for optimal household behavior can be written as $u_Q/u_V = p_J/w$, where p_J is the ideal consumer price index of the consumption basket J. Taking the total differential of this first-order condition, and using Equation (43) and the total differential of the household budget constraint (7) (with $T = 0$), we obtain

$$\tilde{L} = \varepsilon_{LL}^{U} \tilde{w}_R, \tag{44}$$

$$\tilde{C} = \tilde{L} + \tilde{w}_R + (1 - \alpha_C)\sigma_J \tilde{t}_D, \tag{45}$$

$$\tilde{D} = \tilde{L} + \tilde{w}_R - \alpha_C \sigma_J \tilde{t}_D, \tag{46}$$

where $\varepsilon_{LL}^{U} \equiv (1 - L)(\sigma_M - 1)$ stands for the uncompensated wage elasticity of labor supply, and σ_M denotes the substitution elasticity between leisure and composite consumption. $\tilde{w}_R = \tilde{w} - \tilde{p}_J$ represents the relative change in the real after-tax wage, and $\alpha_C \equiv C/wL$ is the share of non-polluting consumption in overall household consumption. The uncompensated wage elasticity is positive if the substitution effect of a change in real wages dominates the income effect, that is, if σ_M exceeds unity. We assume that the labor-supply curve is indeed upward-sloping, as most empirical studies yield positive estimates for this elasticity.

To find the impact of the pollution tax t_R, we use Equations (10) and (11) to log-linearize the demand for the dirty intermediate input conditional on employment:

$$\tilde{R} = \tilde{L} - \varepsilon_R \tilde{t}_R, \tag{47}$$

where $\tilde{t}_R \equiv \mathrm{d}t_R/(1 + t_R)$ and $\varepsilon_R \equiv -[\partial R/\partial t_R][(1 + t_R)/R]$. We can write goods–market equilibrium (39) as (with $t_X = 0$):

$$\omega_L \tilde{L} + \theta_R \omega_R \tilde{R} = (1 - \theta_L)\, \omega_L \left[\alpha_C \tilde{C} + \frac{1 - \alpha_C}{1 + t_D} \tilde{D} \right], \tag{48}$$

[20] In a model with several production factors and several distortionary tax rates, the effect on the *overall* tax base is relevant. In such a model, an expansion in employment (i.e., the base of the labor tax) is neither a necessary nor a sufficient condition for a positive second (non-environmental) dividend.

where $\theta_i \equiv t_i/(1+t_i)$, $i = L, R$, and ω_L and ω_R represent the shares of labor and dirty intermediate inputs in production, respectively. Substituting Equations (44–47) into Equation (48) (with $t_X = 0$), we arrive at

$$\tilde{L} = \frac{\varepsilon_{LL}^U}{\Delta}\left[-\theta_R\omega_R\varepsilon_R\tilde{t}_R - \theta_D\alpha_C(1-\alpha_C)\,\omega_L(1-\theta_L)\,\sigma_J\tilde{t}_D\right], \tag{49}$$

where

$$\Delta \equiv (1-\theta_L)\,\omega_L(1-\theta_D(1-\alpha_C)) - \varepsilon_{LL}^U\left[\omega_L\theta_L + \theta_R\omega_R + \theta_D(1-\alpha_C)\,\omega_L(1-\theta_L)\right] > 0. \tag{50}$$

3.2.1. "Small" environmental taxes. We first consider the equilibrium impacts of an incremental environmental tax, when the initial equilibrium involves no such taxes (i.e. $t_D = t_R = 0$ and hence $\theta_D = \theta_R = 0$). This sets the stage for examining the more general case involving larger pollution taxes. Expression (49) indicates that the introduction of pollution taxes does not affect employment (the expression in square brackets is zero if $\theta_D = \theta_R = 0$), even though the revenues from the pollution taxes allow for lower taxes on labor.

Why is labor supply unaffected? The key to the answer is that environmental taxes are implicit taxes on labor. Like explicit labor taxes, they influence the real-wage and labor-supply incentives. Swapping environmental taxes for labor taxes amounts to substituting implicit labor taxes for the explicit labor tax. While the imposition of the environmental taxes tends to increase labor's tax burden, the reduction in the labor tax tends to reduce it. When the environmental tax is small, these two effects exactly offset each other. Hence the real wage is not changed, which implies that labor supply is unchanged as well[21].

To view this more closely, consider in particular the case where only the tax t_D on the dirty consumer good is raised. Insofar as revenues from the tax t_D can be used to reduce explicit taxes on labor, they raise the real wage. At the same time, however, the tax t_D raises the price of consumption, which has the opposite impact on the real wage. Expression (49) attests to the fact that for a small value of t_D that finances a reduction in t_L, these two effects exactly cancel out. The same result holds in the case where only the tax t_R is introduced. This tax reduces the demand for polluting inputs, thereby reducing labor productivity and thus the before-tax wage. If the initial pollution tax is zero, the adverse effect of the lower before-tax wage on the after-tax wage is exactly offset by the positive effect of lower taxes on labor income[22].

[21] This result depends on leisure being separable from clean and dirty consumption in utility. See Subsection 3.3.2 below for how results may differ under non-separable utility functions.

[22] For a more detailed examination of this issue, see Bovenberg and de Mooij (1994a) and Bovenberg and Goulder (1996).

3.2.2. "Large" environmental taxes. This exact offset does not apply to the case of "large" environmental taxes, however. We can ascertain the impacts of large pollution taxes by analyzing the situation in which environmental taxes are raised from an initial equilibrium in which environmental taxes are positive (i.e. $t_D, t_R > 0$)[23]. Expression (49) indicates that in this case an increase in the pollution tax leads to a reduction in the real wage and a corresponding drop in employment[24].

The negative effect on the real after-tax wage comes about because the lower tax rate on labor income does not fully compensate workers for the adverse effect of the pollution tax on their real after-tax wage. This incomplete offset reflects the fact that environmental taxes tend to be less efficient instruments for raising revenue than a broad-based labor tax. In contrast to a labor tax, pollution taxes on dirty consumption not only affect the labor market but also "distort" the composition of the consumption basket[25]. Furthermore, taxing gross instead of net output by levying pollution taxes on intermediate inputs "distorts" the input mix into production. These "distortions" account for the net reduction in real after-tax income following the revenue-neutral policy change. Of course, these "distortions" in consumption patterns or input choice are desirable on environmental grounds. Indeed, the same features of environmental taxes that make them unattractive from a revenue-raising point of view – their focus on particular inputs or consumption goods – make them attractive as instruments for environmental improvement.

The term in square brackets in Equation (49) represents the additional tax burden associated with a revenue-neutral increase in the pollution tax. This additional burden depends on two elements: the initial levels of pollution taxes, and the substitution elasticities between clean and dirty commodities. The initial pollution taxes regulate the marginal abatement costs. Without prior pollution taxes, reducing a marginal unit of pollution comes free. However, the higher the initial pollution taxes, the larger the marginal costs of increasing environmental quality, since higher initial environmental taxes intensify the adverse revenue effects associated with the erosion of the base from an increment to these taxes. Also, larger substitution elasticities between dirty and clean commodities yield a higher tax burden from a given increment to the pollution tax. Larger substitution elasticities imply larger gross distortions from a given pollution tax (while also implying larger improvements in environmental quality).

3.2.3. Implications for the double-dividend hypothesis and welfare. Having diagnosed the general equilibrium effects on employment, we can now return to Expression (41) to evaluate the double-dividend issue. We have seen that revenue-neutral environmental

[23] The incremental results shown here indicate the impact of a large reform because a large reform's impact is the integral of the impacts of a series of incremental reforms, where the pre-existing pollution tax rates are incrementally larger with each new reform.

[24] Recall that we assume that the uncompensated wage elasticity of labor supply, ε_{LL}^{U}, is positive.

[25] The word "distort" is in quotes to acknowledge the notion that the change in resource allocation may be justified once environmental benefits are taken into account.

tax policies lead to a reduction in employment (for all but infinitesimal environmental tax rates). With a negative value for dL and non-positive values for dD and dR in Equation (41), the tax-base effect is negative. By harming employment, pollution taxes narrow, rather than widen, the tax base. As noted earlier, this means that the non-environmental component of welfare falls. Thus, the double-dividend hypothesis fails.

The absence of the double dividend does not mean that *overall* welfare falls, however. To the contrary, Expression (40) indicates that welfare will rise provided that environmental taxes are not "too large". For "small" environmental taxes, the impact on employment is small. Hence the first right-hand term in Equation (40) is close to zero. At the same time, for small taxes the next two right-hand-side terms in Equation (40) are positive, assuming that marginal environmental benefits ($Nu_Q(-q_{ND})$ and $Nu_Q(-q_R)$) are large for initial reductions in pollution. Thus overall welfare rises[26]. Hence the failure of the double-dividend claim does not imply that green tax reforms are inefficient. It simply means that environmental improvement comes at a (gross) cost.

3.2.4. Significance of second-best considerations. Equations (42) and (49) imply that the gross distortionary cost of a given environmental tax will be larger, the higher the pre-existing taxes on labor. Environmental taxes introduce gross distortions by reducing the labor supply. The larger the pre-existing labor taxes, the greater the wedge between the private and social value of labor, and thus the larger the gross cost associated with a given reduction in the labor supply. Thus, higher pre-existing labor tax rates imply larger costs from given environmental tax reforms.

These results show that partial equilibrium analyses of the gross distortionary costs of environmental taxes can be misleading. Environmental taxes may importantly affect distortions in markets other than those in which the tax is applied. Figure 3 offers the typical partial equilibrium and first-best framework for analyzing welfare effects of an environmentally motivated tax on coal. MC denotes the private marginal costs of producing coal[27]. MC_{soc} represents the social marginal cost curve, incorporating the marginal external damage, MED, from coal combustion. MB stands for the marginal benefit (demand) curve. If a tax is imposed equal to the marginal external damage, social and private marginal costs coincide. The usual textbook analysis regards the welfare gain as area B. This is the value of the environmental improvement ($A+B$) minus the gross costs of the tax (A).

The area A in Figure 3, which corresponds to firm's marginal cost of pollution abatement, can be termed the *primary cost* of the environmental tax. In a world without distortionary taxes, the gross cost of the environmental tax is simply the primary cost.

[26] Recall that t_X is assumed to be zero.

[27] Here we treat coal as equivalent to pollution. In Section 4 we distinguish emissions of polluting compounds from the output or fuel (such as coal) with which pollution is associated. The qualitative results described in the present section are maintained when one refines the analysis to acknowledge this distinction.

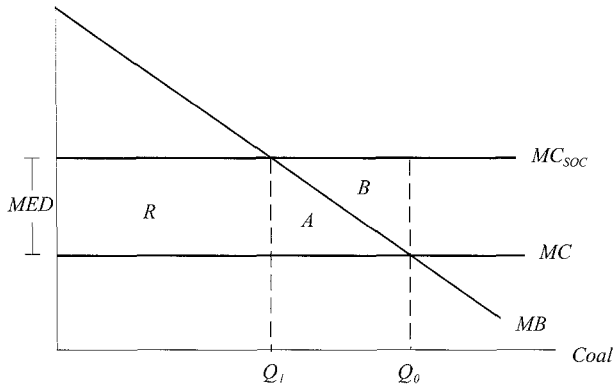


Fig. 3. Typical first-best, partial-equilibrium framework for analyzing efficiency effects of an environmental tax.

However, in the presence of distortionary taxes, the analysis in Figure 3 needs to be modified in two ways. First, the revenues R can be employed to cut distortionary taxes. Such "revenue-recycling" works toward an improvement in efficiency and thus suggests that the partial-equilibrium analysis would overstate the gross cost of environmental taxes. At the same time, since environmental taxes are implicit factor (labor) taxes, a new environmental tax functions as an increase in existing factor taxes. This has the opposite influence on gross costs, tending to raise costs relative to the primary costs indicated in Figure 3 and suggesting that the partial-equilibrium analysis understates gross costs. Under the assumptions in our analysis (where the clean and dirty consumption goods are weakly separable from leisure), the latter effect dominates the former: for a "large" revenue-neutral environmental tax, in the presence of prior taxes the gross costs are higher than the primary costs from Figure 3 – even when revenues are recycled through cuts in the distortionary tax.

Figure 4 schematizes these effects. Adopting terminology similar to that introduced by Parry (1995), we call the former additional effect the *revenue-recycling effect* and the latter additional effect the *tax-interaction effect.* Overall gross cost is primary cost plus the tax-interaction effect minus the revenue-recycling effect. Figure 4 recapitulates the earlier result that an incremental environmental tax reform (associated with incremental abatement) involves zero marginal (gross) cost. However, for a "large" environmental tax (corresponding to a large amount of abatement), the marginal cost is positive. Moreover, under the assumptions in this analysis the tax-interaction effect exceeds the revenue-recycling effect, so that the overall gross cost exceeds the primary cost. Thus, this analysis indicates that in the presence of distortionary taxes the gross

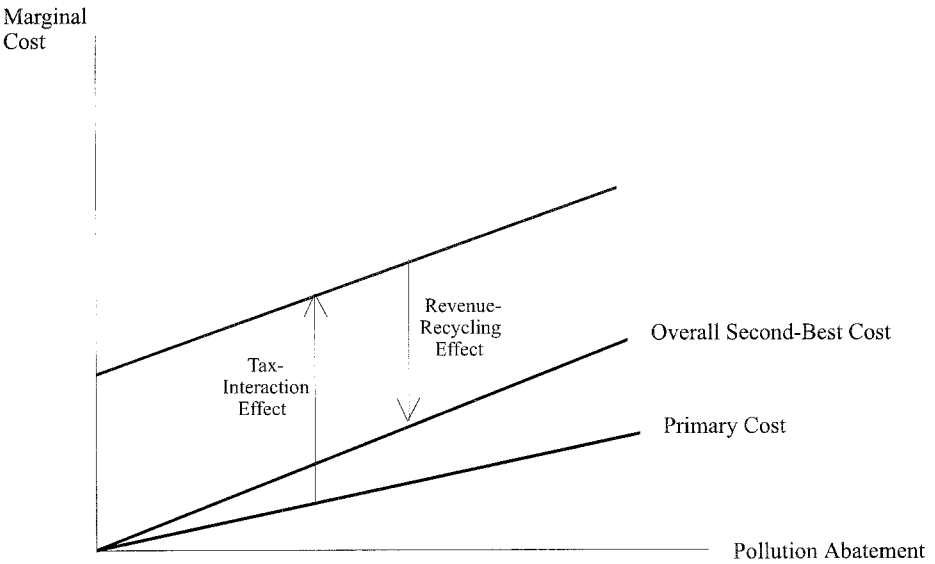


Fig. 4. Marginal costs of pollution abatement in second-best setting.

costs of pollution abatement from revenue-neutral environmental taxes exceed firms' abatement costs[28].

[28] Some distinctions are worth making here. First, the main comparison here is between primary costs and gross costs (of a given environmental tax) within a single, second-best setting. The analysis indicates that, in this setting, the gross costs from a given tax exceed the primary costs of the tax. A different issue is whether, for a given environmental tax, the gross costs in a setting with prior distortionary taxes (a "second-best setting") are greater than the gross costs in a setting without prior distortionary taxes. This would be the case if primary costs were the same in both settings, since gross costs exceed primary costs in the former (second-best) setting, but are equal to primary costs in the latter setting. However, primary costs need not be identical in both settings. Still, numerical simulation studies [see, for example, Goulder, Parry, Williams and Burtraw (1999)] indicate that primary costs are very similar in the presence or absence of distortionary taxes, and that gross costs are indeed higher in a second-best setting than in the absence of prior distortionary taxes. The most policy-relevant comparison seems to be the former one indicating whether, in a realistic second-best setting, gross costs exceed primary costs – that is, whether cost-estimates based on firms' abatement costs understate overall costs.

Second, the focus here is on the costs of a given tax or the costs of given amounts of pollution *abatement*, not the costs of achieving a given amount of environmental *quality*. As pointed out by Gaube (1998) and Metcalf (2000), the underlying level of pollution prior to the introduction of an environmental tax may be different in a world with no prior taxes as compared with a world with distortionary taxes. Pollution levels may be lower in the latter case because of the negative impact of distortionary taxes on factor supply and output. To the extent that initial pollution levels are lower in a second-best setting, the amount of abatement necessary to achieve a given level of environmental quality may be lower in a second-best world. As a result, the cost of achieving a given quality level may be lower in the second-best world, despite the higher costs per unit of abatement.

Note that to obtain a double dividend (negative gross costs), the revenue-recycling effect would have to be large enough to offset both the primary cost and the tax-interaction effect. In Subsection 3.3 below we discuss special circumstances under which this can occur.

In Section 4 we will return to the tax-interaction and revenue-recycling effects, which are important to the choice between taxes and other instruments for environmental improvement.

3.3. Complicating factors

3.3.1. Nature of environmental benefits

Subsection 3.2 assumes that the environment is a public consumption good that enters the utility function in a weakly separable way. As a direct consequence, the improved quality of the environment does not affect the labor market. However, a cleaner environment can affect the labor market through two channels. The first channel applies when environmental quality enters household utility in a non-separable fashion. If environmental quality is complementary to leisure, a cleaner environment makes leisure more enjoyable. In this case, the environmental benefits negatively affect labor supply and thereby magnify the adverse employment effects associated with pollution taxes. If environmental quality is a substitute for leisure, improvements in environmental quality mitigate the adverse employment effects.

The second channel applies when environmental quality exerts a direct effect on labor productivity. To the extent that environmental quality enhances labor productivity, it increases the demand for labor. This offsets the adverse labor-supply impact of the environmental tax and thereby reduces the welfare cost associated with the tax-base erosion effect[29].

These considerations indicate that in principle one should explore the feedback on the economy of a higher supply of the public good of the environment. However, most models exploring the consequences of an environmental tax reform abstract from this feedback. In particular, they ignore the impact of environmental benefits on both labor demand and labor supply. This is a valid assumption only if the environment enters households' utility function as a consumption good in a weakly separable way and is not an input into production.

3.3.2. Inefficiencies in the existing tax system

If the initial tax system is inefficient from a non-environmental point of view, an environmental tax reform may be able to reduce the overall burden of taxation and achieve the double dividend after all. The key requirement is that the revenue-neutral

[29] For a detailed analysis of this issue, see Bovenberg and van der Ploeg (1994a) and Williams (1997).

reform end up alleviating these prior inefficiencies and move the rest of the tax system closer to its non-environmental optimum. This general point can be illustrated with a number of examples.

3.3.2.1. Clean consumption a better substitute for leisure. The first example involves the taxes on clean and dirty consumption in the model of Section 2. Expression (26) shows that the Ramsey tax on dirty consumption should exceed the corresponding tax on clean consumption if, compared to dirty consumption, clean consumption is a better substitute for leisure. Accordingly, if the initial tax system features only a tax on labor (i.e., a uniform tax on clean and dirty consumption), an environmental reform raises private income. Here, raising the tax on dirty consumption and using the revenues to cut the labor tax moves the tax system closer to its optimal Ramsey structure. In this case, the reform boosts employment, thereby alleviating the distortions imposed by the labor tax, and the double dividend thus materializes. However, if dirty consumption is a better substitute for leisure than clean consumption, the Ramsey tax on dirty consumption is negative and the double dividend is even less likely to occur than in the benchmark case examined in Subsections 3.1 and 3.2.

3.3.2.2. Pre-existing subsidies on polluting activities. The overall burden on polluting activities may be too low initially – even from the point of view of maximizing private income – because these activities are subsidized initially. The tax-reform analysis in Section 3.2 illustrates this. If the polluting intermediate input is subsidized in the initial equilibrium (i.e. $t_R < 0$), employment (and hence private income) expands if this subsidy is reduced. Shah and Larsen (1992) emphasize this point in considering the case for carbon taxes in developing countries.

3.3.2.3. Environmental taxes as optimal tariffs. In an open economy, governments can employ pollution taxes as a means of improving the terms of trade. For example, a large oil-importing country may improve its terms of trade if it reduces the demand for oil by raising the tax burden on fossil fuels. Similarly, a large exporting country can boost the prices of its exports by imposing pollution taxes that reduce the supply of these commodities. These terms-of-trade gains shift some of the cost of environmental improvement onto foreigners and lower the domestic welfare cost of environmental policy. If the terms-of-trade gains are large enough, the domestic welfare cost vanishes.

3.3.2.4. Environmental taxes as rent taxes. Environmental taxes may be an implicit way to tax the scarcity rents associated with natural resources. Taxes on the demands for fossil fuels, for example, may be borne largely by the owners of reserves of fossil fuels, as these taxes may reduce significantly the net-of-tax prices of these fuels. To the extent that the burden of environmental taxes falls on the owners of inelastically supplied reserves, the environmental tax functions as a rent tax

and involves no efficiency cost. This improves the prospects for the second (non-environmental) dividend. However, this same phenomenon implies less scope for environmental improvement: the more the tax is borne by owners of reserves, the smaller the increase in the gross-of-tax price to demanders of these fuels. Thus, to advocates of green tax reform, rent taxes are a mixed blessing: they improve the prospects for the second dividend while reducing the scope of the first.

3.3.2.5. Inefficient factor taxation. If the initial tax system involves differences in the marginal excess burdens of various taxes, an environmental tax reform can boost private incomes by shifting the tax burden away from factors with high marginal excess burdens to factors with low marginal excess burdens [see Christiansen (1996), Bovenberg and Goulder (1997), and Goulder (1995a)]. The gross cost of a revenue-neutral environmental tax will be lower to the extent that:

(1) in the initial tax system, the differences in marginal efficiency costs (of various tax instruments) are large,
(2) the burden of the environmental tax falls primarily on the factor with relatively low marginal efficiency cost, and
(3) revenues from the tax are devoted to reducing tax rates on the factor with relatively high marginal-efficiency cost.

These conditions ensure that the efficiency gains from shifting the tax burden from the overtaxed to the undertaxed factor are sufficiently large to offset the costs associated with a cleaner environment.

These considerations may be especially relevant for the mix between capital and labor taxation. Most applied general equilibrium models of the US economy suggest that, compared to taxes on labor income, taxes on capital income tend to produce larger marginal efficiency losses. The most direct way to improve the efficiency of the tax system as a revenue-raising device would be to finance a cut in capital taxes with higher taxes on labor. However, if the government does not want to adopt labor taxes, it can use environmental taxes that are primarily borne by labor[30].

The suboptimality of the initial tax system raises the question why governments have not reformed their tax systems to deal with these inefficiencies. The efficiency rationale for such a tax reform is independent of environmental concerns. However, in some instances, political constraints (perhaps stemming from distributional concerns) may prevent the government from introducing strictly non-environmental tax reforms that enhance the efficiency of the tax system as a revenue-raising device. Under these circumstances, there may be advantages to introducing a package deal in which

[30] See Goulder (1995a) for further discussion of this issue. The welfare effects associated with a sub-optimal initial tax system do not necessarily make environmental tax reforms more attractive, because the burden of the environmental tax could well fall on the factor that is already overtaxed from an efficiency point of view. Indeed, the numerical general-equilibrium analysis in Bovenberg and Goulder (1997) suggests that inefficiencies in the initial US tax system may make carbon taxes less rather than more attractive.

environmental taxes generate revenues that are used to eliminate particularly inefficient taxes. This combination of environmental and non-environmental tax reforms may be necessary to generate sufficient political support for either type of reform. In situations like this, environmental taxes are the lubricating oil that makes possible a tax reform to eliminate particularly "bad" taxes.

3.3.2.6. Inefficient commodity taxation. Some tax policies favor certain forms of consumption over others. For example, the US tax system provides tax deductions for consumer spending on housing or health care. Such policies may be desirable on equity or other grounds, but at the same time they may imply inefficiencies in the allocation of consumer expenditure. To the extent that revenue-neutral environmental tax policies lead to reduced factor tax rates, the values of these tax deductions are reduced. Parry and Bento (2000) show that when this channel is taken into account, the predicted costs of a revenue-neutral environmental tax reform are lowered significantly. In particular, the gross costs of such a reform might well be significantly below the primary costs.

3.3.3. Involuntary unemployment

The previous analysis assumed a well-functioning labor market with a flexible wage rate that assures full employment. In Europe, where involuntary unemployment is widespread, there has been considerable interest in green tax reform as a vehicle for reducing unemployment as well as improving the environment. This has prompted several studies investigating the effects of revenue-neutral environmental tax reforms in situations involving involuntary unemployment.

Bovenberg and van der Ploeg (1998a) analyze the consequences of an environmental tax reform in a model where involuntary unemployment stems from a rigid consumer wage. Hence this model incorporates labor-market distortions as well as the tax distortions already considered. In addition to labor, a clean non-labor production factor, which is fixed in supply, enters production[31]. Non-labor income is subject to a fixed tax rate of less than 100 percent.

The analysis shows that an environmental tax reform may reduce involuntary unemployment by expanding labor demand. Specifically, in the presence of a non-labor production factor, an environmental tax reform can shift part of the tax burden away from labor to the inelastically supplied non-labor factor. This tax-shifting effect exerts a positive impact on employment because it allows for a fall in wage costs, thereby boosting labor demand. In this model, there is a wedge between the marginal social value and marginal social cost of employment both because of the distortionary labor tax and because of the gap between the actual consumer wage and the reservation wage (i.e. the wage at which households would be willing to work). Hence the increase in employment from the environmental tax reform yields a first-order welfare gain.

[31] Without this latter production factor, both the consumer wage and the production wage would be fixed, and the market wage would be overdetermined.

These results are another example of how the prospects for the double dividend improve when the initial tax system is inefficient from a non-environmental point of view (see also Subsection 3.3.2). Since the non-labor production factor is fixed in supply, a 100 percent tax on non-labor income would be most efficient. However, if such a tax is infeasible for political or other reasons, a second-best alternative is to introduce the pollution tax, an implicit tax on the fixed factor (and labor). Substituting the pollution tax for the labor tax improves efficiency by shifting more of the burden of taxation onto the fixed factor.

Several papers [see, e.g., Koskela and Schöb (1999), Nielsen, Pedersen and Sørensen (1995), Schneider (1997) and Bovenberg and van der Ploeg (1998b)] explore how a green tax reform affects equilibrium unemployment in models with endogenous wage-setting. In these models, the impact on employment depends mainly on how an environmental tax reform affects unemployment benefits, which determine the threat point of employees in the bargaining process between employers and employees. In particular, if unemployment benefits are a fixed proportion of income in employment (implying a fixed replacement ratio), then all taxes are completely borne by employees [see Layard, Nickell and Jackman (1991)]. Hence, an environmental tax reform affects neither wage costs nor unemployment in equilibrium. The importance of the benefit regime applies to most equilibrium models of unemployment, including models of union–firm bargaining, monopoly unions, efficiency wages, and job search.

Koskela and Schöb (1999) illustrate these principles in the context of a model of wage bargaining between unions and employers. They show that the employment effects of an environmental tax reform involving pollution taxes on dirty consumption depend crucially on the taxation of unemployment benefits. In particular, employment may expand if unemployment benefits are neither subject to the labor-income tax nor indexed to the consumer price index. In that case, the unemployed pay the higher pollution taxes on consumption but are compensated neither by lower taxes on labor income nor by higher gross benefits. Indeed, whereas the pollution tax hits workers and unemployed alike, only workers benefit from the recycled revenues in the form of lower taxes on labor. In this way, the environmental tax reform shifts the tax burden away from workers towards the unemployed. This tax-shifting effect makes the outside option of unemployment less attractive for workers, thereby moderating wage costs and thus boosting labor demand.

3.4. Numerical assessments of a green tax reform

The interest in green tax reform has prompted several empirical studies of potential reforms. Many of these studies employ sophisticated numerical general-equilibrium models that contain considerably more detail than the analytically tractable model developed above.

3.4.1. Impacts on consumption and welfare

Table 4 summarizes results from numerical studies of a potential reform that has gained especially great interest: a revenue-neutral carbon tax policy. The

Table 4
Numerical assessments of welfare impacts of revenue-neutral environmental tax reforms

Model	Reference	Country	Type of environmental tax	Method of revenue replacement	Welfare effect
DRI	Shackleton et al. (1996)	USA	Phased-in carbon tax[a]	Personal tax cut	–0.39[b]
Goulder	Goulder (1995b)	USA	$25/ton carbon tax	Personal tax cut	–0.33[c]
	Goulder (1994)	USA	Fossil fuel Btu tax	Personal tax cut	–0.28[c]
Jorgenson–Wilcoxen	Shackleton et al. (1996)	USA	Phased-in carbon tax[a]	Capital tax cut	0.19[d]
LINK	Shackleton et al. (1996)	USA	Phased-in carbon tax[a]	Personal tax cut	–0.51[b]
Proost–van Regemorter	Proost and van Regemorter (1995)	Belgium	Hybrid of carbon and energy tax	Payroll (social security) tax cut	–3.45[d]
Shah–Larsen	Shah and Larsen (1992)	USA	$10/ton	Personal tax cut	–1049[e]
		India			–129
		Indonesia			–4
		Japan			–269
		Pakistan			–23

[a] Beginning at $15/ton in 1990 (period 1), growing at five percent annually to $39.80 per ton in 2010 (period 21), and remaining at that level thereafter.
[b] Percentage change in the present value of consumption; the model does not allow for utility-based welfare measures.
[c] Welfare cost per dollar of tax revenue, as measured by the equivalent variation.
[d] Equivalent variation as a percentage of benchmark private wealth.
[e] Compensating variation in levels (millions of US dollars).

table presents results from seven numerical models. These are the Goulder and Jorgenson–Wilcoxen intertemporal general-equilibrium models of the USA, the Proost–van Regemorter general-equilibrium model of Belgium, the DRI and LINK econometric macroeconomic models of the USA, and the Shah–Larsen partial equilibrium model, which has been applied to five countries, including the USA[32]. The results in Table 4 are for the revenue-neutral combination of an environmental tax (usually a carbon tax) and reduction in the personal-income tax, except in cases where this combination was not available.

All welfare changes abstract from changes in welfare associated with improvements in environmental quality (reductions in greenhouse-gas emissions). Thus they correspond to the gross distortionary cost concept discussed above. In the Goulder, Jorgenson–Wilcoxen, and Proost–van Regemorter models, welfare changes are reported in terms of the equivalent variation; in the Shah–Larsen model, the changes are based on the compensating variation[33]. In the DRI and LINK macroeconomic models, the percentage change in aggregate real consumption substitutes for a utility-based welfare measure[34].

In most cases, the revenue-neutral green tax swap involves a reduction in welfare, that is, entails positive gross costs. This militates against the double-dividend claim. Results from the Jorgenson–Wilcoxen model, however, support the double-dividend notion. Relatively high interest elasticities of savings (a high capital-supply elasticity) and the assumption of perfect capital mobility across sectors may partially explain this result, at least in the case where revenues from the carbon tax are devoted to cuts in marginal taxes on capital. These assumptions yield large marginal excess burdens (MEB's) from taxes on capital, considerably larger than the MEB's from labor taxes.

[32] For a more detailed description of these models, see Goulder (1995b), Jorgenson and Wilcoxen (1990, 1996), Shackleton et al. (1996), Proost and van Regemorter (1995) and Shah and Larsen (1992). In the models with explicit utility functions (all except the LINK and DRI models), environmental quality is implicitly regarded as separable in utility from leisure and consumption. These models also assume, as in the analysis of Subsections 3.1 and 3.2, that leisure and consumption are weakly separable. These assumptions reflect the lack of empirical information about the relative substitutability of consumer goods with leisure. Under these circumstances, assuming that all goods are equal substitutes with leisure (weak separability) seems reasonable.

The Shah–Larsen model is the simplest of the models, in part because it takes pre-tax factor prices as given. Despite its simplicity, the model addresses interactions between commodity and factor markets and thus incorporates some of the major efficiency connections discussed earlier.

[33] The equivalent variation is the lump-sum change in wealth which, under the "business-as-usual" or base case, would leave the household as well off as in the policy-change case. Thus a positive equivalent variation indicates that the policy is welfare-improving. The compensating variation is the lump-sum change in wealth which, in the policy-change scenario, would cause the household to be as well off as in the base case. In reporting the Shah–Larsen results we adopt the convention of multiplying the compensating variation by –1, so that a positive number in the table signifies a welfare improvement here as well.

[34] The demand functions in these models are not derived from an explicit utility function. Hence they do not yield utility-based measures.

As indicated in Subsection 3.3.2, if the MEB on capital significantly exceeds that on labor, and the environmental reform shifts the tax burden on to labor, the double dividend can arise. Thus, the large MEB's from capital taxes help explain why, in the Jorgenson–Wilcoxen model, a revenue-neutral combination of carbon tax and reduction in capital taxes involves negative gross costs, that is, produces a double dividend.

Identifying the sources of differences in results across models is difficult, in large part because of the lack of relevant information on simulation outcomes and parameters. Relatively few studies have performed the type of analysis that exposes the channels underlying the overall impacts. There is a need for more systematic sensitivity analysis, as well as closer investigations of how structural aspects of tax policies (type of tax base, narrowness of tax base, uniformity of tax rates, etc.) influence the outcomes. In addition, key behavioral parameters need to be reported. Serious attention to these issues will help explain differences in results and, one hopes, lead to a greater consensus on likely policy impacts.

3.4.2. An employment dividend?

In Europe, policy makers have been especially interested in the possibility that green tax reforms could raise employment. Many politicians have supported reforms in which pollution taxes would be introduced and the revenues devoted to cuts in labor taxes. The preoccupation with employment impacts reflects in part the relatively high rates of unemployment prevailing in many European countries.

In models with only labor as a primary factor of production, the employment impacts of a revenue-neutral environmental tax reform are directly related to the impacts on the non-environmental component of welfare. In Expression (41), the sign of dL determined the sign of the non-environmental component of welfare. In more detailed models – in particular, models that distinguish more than one primary factor of production – the employment dividend is no longer tied so closely to the non-environmental-welfare dividend. Revenue-neutral reforms can produce an increase in employment without raising real incomes and non-environment-related welfare.

The crucial requirement for an increase in employment is that the reforms shift the tax burden from labor to other primary factors. Specifically, in models with capital and labor, the prospects for an employment dividend are enhanced to the extent that:

(1) The industry or industries on which the environmental tax is levied (that is, the pollution-intensive industries) feature a relatively low labor intensity in comparison with other industries.
(2) Revenues from the revenue-neutral policy are devoted primarily to cuts in labor taxes (rather than taxes on capital)[35].

[35] A further consideration is whether the environmental reform produces a *tax-shifting effect* that has a positive efficiency impact (see discussion in Subsection 3.3.2). In particular, if the prior tax system overtaxes labor relative to capital, and the environmental reform shifts the burden from labor to capital,

Many numerical models have examined the employment impacts of revenue-neutral reforms, considering a wide range of energy and environmental taxes. Some models incorporate considerable detail on labor markets, including wage formation by unions and labor-market wage rigidities that lead to involuntary unemployment[36]. Employment impacts can be quite sensitive to the specification of these features of the labor market. Although results vary widely, they indicate that the employment dividend can materialize when revenues are recycled through cuts in labor taxes and when the industries facing the environmental tax are not exceptionally labor-intensive[37]. An employment dividend also can arise if the revenue-neutral reform tends to shift the burden of taxation from labor to transfer recipients. This may occur, for example, when the government introduces an environmental tax and devotes the revenues to cuts in the labor tax. The environmental tax raises the real cost of output, but labor enjoys a reduction in the labor tax that more than offsets this increase. Transfer recipients, in contrast, are not compensated for the reduction in the real value of their transfers. Under these circumstances, labor enjoys an increase in real income from the revenue-neutral reform, despite the overall gross cost of the reform, because the tax burden is shifted to transfer recipients. Consequently, employment rises. Results from the MIMIC numerical general equilibrium model of the Netherlands exhibit this phenomenon.

4. Alternatives to pollution taxes

While economists tend to favor taxes as instruments for environmental protection, most environmental regulation is accomplished through other instruments. There are only a few instances of "environmental"[38] taxes in the USA – a tax on gasoline, on motor fuels, on oil spills, on ozone-depleting chemicals, and on chemical feedstocks (associated with toxic-waste production) – and the bulk of environmental regulation is accomplished through mandated technologies or performance standards. In other countries the emphasis on taxes is even lighter in comparison with other instruments. In this section we consider some important alternatives to taxes.

4.1. Instrument choice in a certainty context

We compare different instruments using a slightly altered version of the model used in Section 2. Here we abstract from intermediate inputs and focus on pollution

tax-shifting will lead to greater efficiency in the relative taxation of labor and capital. This, in turn, helps to raise the real wage, which tends to boost employment.

[36] See, for example, Brunello (1996) and Capros et al. (1996).

[37] See, for example, the results from the collection of models examined in Carraro and Siniscalco (1996).

[38] The political motivation for introducing these taxes need not be environmental. Hahn (1989), Fullerton (1996) and Stavins (2000) examine the various environmental taxes employed in the USA and discuss their original rationales.

associated with the production or use of the "dirty" consumption good, D. Labor is the only input into production. To facilitate comparisons of alternative instruments, we generalize slightly the relationship between environmental quality, Q, and the output of the dirty consumption good, D. In particular, we now include in the model pollution abatement, A. By devoting resources to pollution abatement (for example, by implementing new, cleaner production processes), firms can reduce the amount of pollution per unit of production of the dirty consumption good. In this setting, environmental quality depends on pollution emissions E, which in turn depend on the level of production of the dirty consumption good and on the level of abatement expenditure. Hence $Q = q(E(ND, A))$, with $\partial Q/\partial E < 0$, $\partial E/\partial ND > 0$, $\partial E/\partial A < 0$.

In this altered model, the economy's transformation surface is:

$$NL = G + NC + ND + A, \tag{51}$$

where, as before, we normalize units so that marginal rates of transformation are unity.

To provide a reference point, we first derive the welfare impacts of a tax t_E on emissions of pollution. Taking the derivative of the representative household's utility with respect to t_E yields:

$$\frac{\mathrm{d}U/\mathrm{d}t_E}{u_C} = \frac{1}{N}(t_E - t_E^P)\frac{\partial E}{\partial t_E} + \eta t_L w \frac{\partial L}{\partial t_E} + \frac{1}{N}(\eta - 1)\left[\frac{\partial (Et_E)}{\partial t_E}\right]. \tag{52}$$

The welfare impact of an incremental change in t_E has three components. The first term on the right-hand side of Equation (52) is the *primary gain,* the welfare change associated with a change in environmental quality, net of the primary cost (private abatement cost). This term is positive if the pollution tax is below the Pigouvian rate, $t_E^P = -Nu_Q q_E/u_C$. This term vanishes if the pollution tax t_E equals the Pigouvian tax t_E^P, that is, if the external effects of pollution are fully internalized. The second term represents the *tax-interaction effect* introduced in Section 3 [see Parry (1995, 1997)]. As discussed earlier, because the environmental tax raises the costs of production and the prices of goods in general, it acts as an implicit tax on labor. The tax-interaction effect is the adverse welfare impact that results from this implicit tax's negative impact on the real wage and employment. The third term on the right-hand side is the *revenue-recycling effect.* It represents the beneficial welfare impact stemming from the environmental tax's generation of revenues that can be used to finance cuts in the labor tax. Two conditions ensure that the revenue-recycling effect is indeed positive (at the margin). First, the MCPF must exceed one. Second, the slope of the Laffer curve for the pollution tax must be increasing (i.e., $\partial(Et_E)/\partial t_E > 0$).

4.1.1. Pollution quotas

Now compare the impact of the pollution tax with that of a pollution quota. Under the quota policy, the government gives out free to each polluting firm a fixed number of

pollution permits, where each permit entitles the owner to a given amount of emissions of pollution[39]. For now we treat firms as identical and as receiving the same quotas or numbers of permits. Thus, if the government's targeted level of aggregate emissions is $\bar{E}$, then $\bar{E}/K$ is the quota allocation to each of the K polluting firms. Under these conditions there is no scope for gains from trading permits. In Subsection 4.1.2 we will consider permits trades among heterogeneous firms.

Taking the derivative of utility with respect to $\bar{E}$ yields

$$\frac{\mathrm{d}U/\mathrm{d}\bar{E}}{u_c} = \frac{1}{N}(\bar{t}_E - t_E^P) + \eta t_L w \frac{\partial L}{\partial \bar{E}}. \tag{53}$$

Let $\bar{t}_E$ represent the *virtual* tax rate on emissions associated with $\bar{E}$, that is, the emissions tax rate (with lump-sum replacement of tax revenues) that would yield the level of emissions under the quota. Multiplying Equation (53) by $\partial\bar{E}/\partial\bar{t}_E$ yields an expression very similar to (52) except that the revenue-recycling effect – the far-right term in Equation (52) – is missing. Since the pollution quota does not raise revenue[40], it cannot finance cuts in the pre-existing labor tax. Thus it cannot yield the beneficial welfare impact associated with recycling of environmental tax revenues.

The absence of the revenue-recycling effect is a disadvantage of quotas relative to emissions taxes. The cost of achieving a given amount of pollution abatement is higher under quotas than under emissions taxes – assuming that revenues from the emissions taxes are used to finance cuts in pre-existing distortionary taxes. The significance of the revenue-recycling effect is most easily seen if one considers the welfare cost of the first unit of pollution abatement (that is, the impact of raising t_E or $\bar{t}_E$ from an initial value of zero). Under the conditions on utility in Subsection 3.2, this increment to the pollution tax t_E produces a revenue-recycling effect that exactly offsets the tax-interaction effect: thus the last two terms in Equation (52) cancel out. Hence, so long as marginal environmental benefits are strictly positive (i.e., $t_E^P > 0$), incremental pollution abatement by way of the pollution tax raises welfare. This result is consistent with traditional first-best analyses. Under the pollution quota, however, positive marginal environmental benefits do not guarantee a welfare improvement from incremental abatement. Under this policy the first unit of abatement produces a strictly positive tax-interaction effect, and there is no revenue-recycling effect to offset it. Expression (53) shows that welfare rises only if the marginal environmental benefits

[39] The instrument analyzed here is sometimes referred to as a uniform performance standard. However, we apply the term "performance standard" below to an instrument that limits the emissions rate rather than the quantity of emissions.

[40] To the extent that the quota affects production costs and thereby alters labor supply, it will affect the labor tax base and tax revenues. This effect is captured in the far-right (tax-interaction effect) term in Equation (53).

are large enough to overcome this tax-interaction effect. Specifically, a welfare gain requires that

$$\frac{1}{N}t_E^P > \eta t_L w \frac{\partial L}{\partial \bar{E}}. \tag{54}$$

Thus, in a second-best setting, the choice between a pollution tax and pollution quota can affect not only the level but also the sign of the welfare impact!

In Subsection 2.6.3 we observed that when revenues from a carbon tax are returned lump-sum, environmental benefits must exceed a certain threshold value for such pollution taxes to yield an efficiency improvement. Similarly, under non-auctioned pollution quotas the absence of the revenue-recycling effect implies that environmental benefits must also exceed a threshold value before efficiency gains become possible.

The efficiency advantage of the emissions tax over the emissions quota is premised on the idea that revenues from the emissions tax are used to finance cuts in pre-existing distortionary taxes. If, instead, the revenues from the emissions tax were returned in lump-sum fashion, the revenue-recycling effect would disappear and this efficiency advantage would vanish. Conversely, the disadvantage of the emissions quota stems from the fact that the quotas are not auctioned, yield no revenues, and thus do not exploit the revenue-recycling effect. Auctioned quotas whose revenues finance cuts in distortionary taxes would suffer no disadvantage relative to the emissions tax considered here. Thus, what is crucial to the efficiency impact is whether the policy manages to counter the tax-interaction effect by exploiting the revenue-recycling effect.

There is another way to interpret the parallel results under (non-auctioned) emissions quotas and under emissions taxes with lump-sum replacement of the revenues. Under both of these policies, the government effectuates a lump-sum transfer to households, either explicitly or by generating untaxed quota-related rents. In a second-best setting, such transfers are costly because they ultimately must be financed through distortionary taxes[41]. Hence the costs of achieving given emissions reductions through pollution taxes with lump-sum replacement, or through non-auctioned pollution quotas, are greater than the costs under a pollution tax (or auctioned quota) with revenues devoted to cuts in the marginal rates of pre-existing distortionary taxes.

For given levels of abatement, the efficiency advantage of pollution taxes (with revenues devoted to cuts in marginal tax rates) over non-auctioned quotas rises with the size of the pre-existing tax rate on labor. This occurs because a higher pre-existing tax rate implies a larger revenue-recycling effect.

Goulder, Parry and Burtraw (1997) show that the efficiency advantage of taxes over (non-auctioned) quotas declines with the extent of pollution abatement. In the limiting

[41] Fullerton and Metcalf (2001) emphasize the importance of policy-induced rents in analyzing the different efficiency costs of incremental pollution taxes, quotas, and technology restrictions.

case of 100 percent pollution abatement (either through a prohibitively high pollution tax or a "quota" of zero) the two policies generate identical efficiency impacts. In this extreme case, neither the pollution tax nor the quota raises any revenue: hence the revenue-recycling effect is absent under both policies and the two policies generate the same outcome. More generally, the *marginal* revenue [$\partial(Et_E)/\partial t_E$ in Equation (52)] generated by a pollution tax usually declines and eventually becomes negative as the pollution tax (or amount of abatement) becomes quite large. At the point where the marginal revenue becomes negative (that is, where the peak of the Laffer curve is reached), the marginal revenue-recycling effect from the pollution tax switches sign – an increment to the pollution tax reduces tax revenues and thus necessitates an *increase* in the tax rate on labor. Because the *marginal* revenue-recycling effect declines, the total gross costs of pollution abatement increase more rapidly, as a function of abatement, under the pollution tax than under the quota. At 100 percent abatement the total gross costs become identical[42].

Goulder, Parry and Burtraw (1997) have examined these issues in the context of the regulation of sulfur-dioxide emissions from US coal-fired electric power plants[43]. Title IV of the 1990 Clean Air Act Amendments restricts emissions of SO_2 through a system of freely offered (or "grandfathered") emissions permits, which have similar efficiency properties to quotas. An alternative regulatory approach would be to auction the emissions permits or, equivalently, to impose an emissions tax. Figure 5 indicates Goulder, Parry and Burtraw's estimates of the costs of SO_2 emissions reductions under freely offered permits (actual policy) and under auctioned permits. These estimates stem from a simple numerical general-equilibrium model. The two solid lines in the figure are the ratios of total costs in a second-best setting (with a positive pre-existing tax rate on labor equal to 0.4) to total costs in a first-best setting (with no pre-existing tax on labor). In the case of auctioned permits (or pollution taxes), the line is almost perfectly horizontal: this ratio is approximately constant throughout the entire range of possible emissions reductions (0 to 20 million tons). Second-best considerations raise the costs of auctioned permits by about 30 percent, regardless of the extent of emissions abatement. In contrast, for the actual policy of freely offered emissions permits, the ratio of total cost is very sensitive to the extent of abatement. Under

[42] Thus the *marginal* costs of abatement eventually become higher under the pollution tax than under the quota. This implies that if environmental damages are sufficiently large to justify a large amount of pollution abatement, the optimal amount of pollution abatement will be higher under the quota than under the pollution tax. Somewhat less abatement is justified under the pollution tax because further abatement involves a costly loss of revenue. Indeed, the early literature on the revenue-raising capacity of pollution taxes focussed on the sign of the revenue-recycling effect by computing the revenue-maximizing rate in a partial-equilibrium setting. See Terkla (1984) and Lee and Misiolek (1986). Lee and Misiolek compute elasticities of the pollution tax base to investigate whether current pollution taxes are set to maximize revenue.

[43] Parry, Williams and Goulder (1999) apply similar models to compare the costs of carbon-dioxide abatement under grandfathered (freely offered) carbon quotas and carbon taxes (or auctioned quotas).

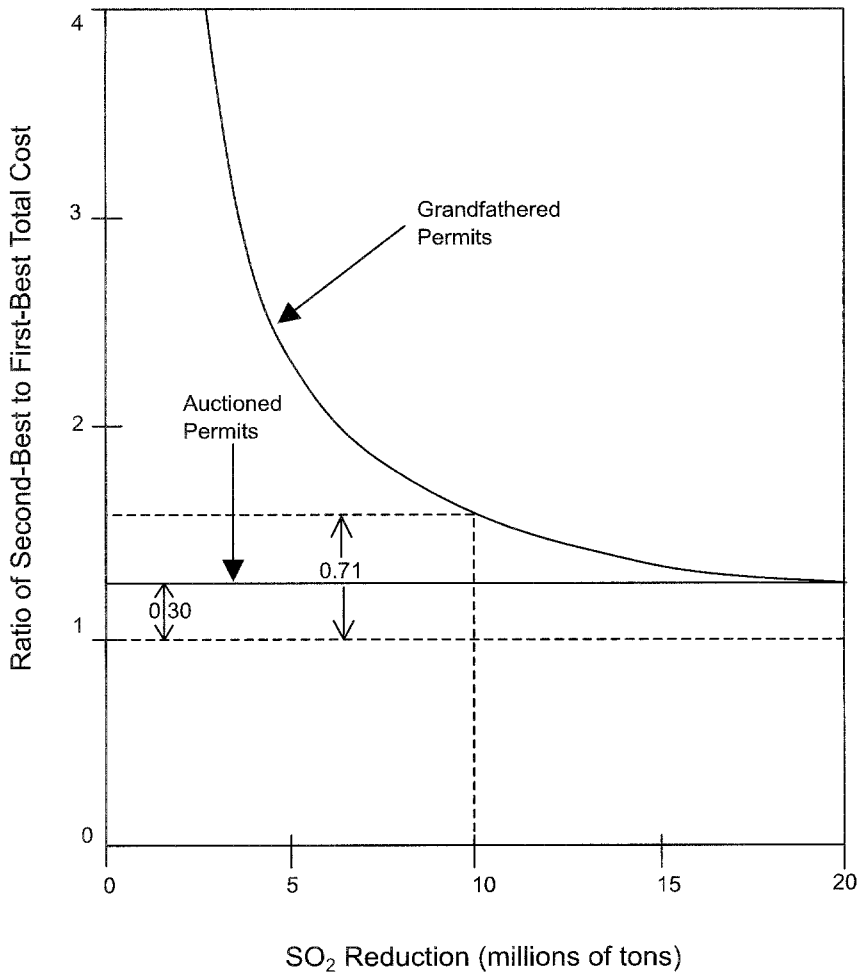


Fig. 5. Total costs of auctioned permits (emissions tax) and grandfathered permits in a second-best setting.

this policy the ratio begins at infinity[44], but as the level of abatement approaches 100 percent, the ratio of total costs approaches the ratio for auctioned permits. The 1990 Clean Air Act Amendments call for a 10-million-ton (or approximately 50 percent) reduction in SO_2 emissions. Significant distributional or political objectives may be served by grandfathering (i.e., giving permits out free to existing firms), but Figure 5's results indicate that they come at a high price in terms of the social cost of

[44] This is in keeping with the fact that the intercept of the marginal cost function is positive for this policy in a second-best world and zero in a first-best world.

abatement. At 10 million tons of abatement, annual total costs under the actual policy are estimated to be 71 percent (or $907 million) higher than they would be in a first-best world. As indicated in this figure, more than half of this additional cost could be avoided by auctioning the permits or employing an SO_2 tax. These results indicate that pre-existing taxes and the presence or absence of revenue-recycling exert a substantial impact on the costs of environmental policies.

4.1.2. Tradeable emissions permits

Until now we have abstracted from the considerable heterogeneity among producers within a given industry. In fact there is considerable heterogeneity of this sort, and this poses significant regulatory challenges. We have just noted that non-auctioned emissions quotas suffer a disadvantage relative to taxes because they fail to generate revenues and thus cannot exploit the revenue-recycling effect. In the presence of heterogeneity, another disadvantage may arise, depending on whether the quotas are tradeable. Efficient pollution regulation requires that marginal costs of abatement be equal across sources. If regulators impose non-tradeable pollution quotas, they are unlikely to have sufficient information to impose such quotas in a way that succeeds in equating the firms' marginal abatement costs. In contrast, the imposition of a pollution tax (or of tradeable pollution quotas) encourages firms to equate their marginal costs of abatement to the value of the pollution tax. Thus, if firms within the polluting industry face the same tax, the tax will promote equality of marginal abatement costs.

Despite this potential efficiency advantage of taxes, regulators are often reluctant to introduce them, in part because of political opposition connected with the fact that taxes require firms to pay for each unit of emissions, while quotas do not. A system of tradeable emissions permits may offer a partial solution to this dilemma. Such a system has the potential to yield a cost-effective allocation of abatement effort (equality of marginal abatement costs across firms) while, like quotas, enabling firms to produce a certain amount of emissions without being charged for it.

Tradeable emissions permits systems were first described in theoretical terms by Crocker (1966), Dales (1968), and Montgomery (1972). Subsequently, Hahn and Noll (1982), Tietenberg (1985), and others have shown how such systems could be implemented in realistic regulatory contexts. Under such a system, firms are allocated permits entitling them to certain levels of emissions over a given period of time. Firms can trade these permits and thereby either augment or reduce their emissions entitlements. In theory, permits trades lead to an equilibrium in which marginal costs of pollution abatement are equalized across firms – thus the equilibrium achieves given overall abatement targets cost-effectively[45].

[45] In practice, the efficiency of tradeable-permits markets may be compromised by non-competitive market conditions [see Hahn (1984) and Misiolek and Elder (1989)] and transactions costs [see Stavins (1995)].

The basic workings of a tradeable-permits system are as follows. Suppose that regulators issue to firms a total of Z pollution permits, where each permit entitles its owner to one unit of pollution emissions (over a given interval of time). Let z_{0j} represent the number of permits initially allocated (free) to firm j. Firms decide to purchase additional permits or sell some of their permits in order to minimize their costs. Let e_{0j} represent the firm's "unconstrained" emissions level, that is, the amount of emissions that the firm would generate if there were no regulation, and let e_j denote the firm's chosen emissions level after the implementation of the permits market.

The firm's problem is choose the emissions level e_j to

$$\min_{e_j} c(e_{0j} - e_j) + p(e_j - z_{0j}), \tag{55}$$

where $e_{0j} - e_j$ represents the firm's level of emissions abatement, $c(\cdot)$ is the firm's abatement cost function, and p is the market price of permits. The second term in Equation (55) indicates that to be entitled to generate emissions in excess of the amount z_{0j}, the firm must purchase the additional permits $e_j - z_{0j}$; similarly, if the firm wishes to reduce its emissions below z_{0j}, it can sell its excess permits $z_{0j} - e_j$. The first-order condition for this problem is

$$-\frac{\partial c}{\partial e_j} = p. \tag{56}$$

Firms purchase or sell permits until the marginal cost of abatement equals the price of a permit. A firm whose current stock of permits implies lower marginal costs of abatement than p will wish to sell permits (and be compelled to abate more); a firm whose current stock implies higher abatement costs than p will wish to purchase more permits. The market price of permits adjusts to a level that clears the permits market. In equilibrium, marginal abatement costs of all firms are equated to the market price, p. Thus, purchases and sales of permits generate production efficiency[46].

Tradeable-permits systems are a hybrid of quantity- and price-based regulations. They are quantity-based in that the total acceptable amount of emissions is set by the regulatory authority (in the choice of Z). They are price-based in that market forces determine the equilibrium prices of permits and the ultimate allocation of permits across firms. Because they help bring marginal abatement costs into alignment, they tend to be able to achieve given pollution-reduction targets at lower cost than would be possible under systems of mandated (non-tradeable) emissions quotas. We mentioned that a tradeable-permits system has been implemented in the USA as

[46] By achieving production efficiency, tradeable-permits systems tend to imply lower output prices relative to a system of non-tradeable quotas that achieves the same aggregate emissions reductions. This means that the tax-interaction effect will be smaller under tradeable permits than under non-tradeable quotas, which augments the efficiency advantage of the permits approach.

part of the regulation of SO_2 emissions from coal-fired electric power plants under Title IV of the 1990 Clean Air Act Amendments. Tradeable-permits programs have also been introduced in the USA to control the lead content of gasoline from petroleum refineries, to reach compliance with the Montreal Protocol's mandated reductions in the production of chlorofluorocarbons (which contribute to the greenhouse effect), and to control emissions of sulfur oxide and nitrous oxide compounds from stationary sources in the Los Angeles airshed[47]. The Los Angeles program is estimated to yield cost savings of 40–50 percent over the period 1995–2010 relative to a system in which the same aggregate emissions reductions were achieved in the absence of trades.

It should be noted that the adoption of a tradeable-permits system may help foster an efficient allocation of abatement effort, but does not generally guarantee an efficient level of aggregate pollution. The latter requires that the number of permits (Z) be chosen optimally. In addition, systems in which permits are initially freely allocated (or grandfathered) are at a disadvantage relative to emissions taxes in terms of efficiency. Such systems share the drawback of non-tradeable pollution quotas in that they fail to exploit the revenue-recycling effect. An alternative regulatory approach would be for the government to auction the permits. This alternative approach is formally equivalent to introducing an emissions tax. Like an emissions tax, this approach exploits the revenue-recycling effect and involves a smaller efficiency cost than grandfathered permits.

4.1.3. Subsidies to pollution abatement

Another way to discourage pollution emissions is to subsidize the abatement of pollution. In the case of an abatement subsidy, the government effectively grants pollution rights to firms, and obligates taxpayers to compensate firms for any reductions in pollution. This is consistent with the *victim pays* principle whereby the recipients of pollution must pay to induce pollution reductions. In contrast, an emissions tax effectively grants potential pollutees (taxpayers) the right to a pollution-free environment, and obligates firms to pay taxpayers (by paying emissions taxes) for the privilege of violating this right. This is consistent with the *polluter pays* principle whereby the generators of pollution must pay for the privilege of polluting[48].

[47] See Stavins (2000) and Tietenberg (1997) for a review of permits trading programs in the USA.

[48] The victim-pays and polluter-pays principles represent differing initial specifications of property rights. Under the victim-pays principle, society initially offers potential victims the right to be free of pollution, and polluters must pay victims for the privilege of violating that right. In contrast, under the polluter-pays principle, potential polluters initially enjoy the right to pollute, and victims must compensate polluters if they wish to be free of pollution. These contrasting initial specifications parallel the different cases considered by Coase (1960) in his famous theorem concerning the possibilities for efficiently solving externalities problems through voluntary agreements by affected parties. Coase's theorem asserts that an efficient outcome can be produced under either specification of initial property rights. However, as we will discuss below, when one moves from the Coasean setting involving voluntary arrangements

Consider the case where the government rewards firms at the rate s for each unit of pollution abatement relative to some baseline amount e_0. Thus the firm with emissions e receives the subsidy payment $s(e_0 - e)$. Under these circumstances the firm loses the value s for each positive increment of pollution; hence, at the margin, the cost of emissions is the same as that of a pollution tax at the rate s.

Assume that, prior to regulation, K identical firms were responsible for pollution, with each firm generating emissions equal to e_0. Using the same approach as was employed with emissions taxes, we obtain the following expression for the welfare impact of the abatement subsidy:

$$\frac{\mathrm{d}U/\mathrm{d}s}{u_C} = \frac{1}{N}(s - t_E^P)\frac{\partial E}{\partial s} + \eta t_L w \frac{\partial L}{\partial s} + \frac{1}{N}(\eta - 1)\left[\frac{\partial[(E - E_0)s]}{\partial s}\right]. \tag{57}$$

This expression differs from Equation (52) only in that the subsidy s replaces the emissions tax t_E and $E - E_0$ replaces E in the far-right term representing the revenue-recycling effect. Thus, the subsidy produces the same primary gain and tax-interaction effects as does the emissions tax. This is consistent with the notion that, for a firm with baseline emissions of e_0, an emissions subsidy at rate t is equivalent to an emissions tax of that rate on emissions e plus a lump-sum payment of te_0. The revenue-recycling effect is different under the subsidy, however. Under the subsidy, the revenue-recycling effect works against efficiency, since the government must now raise labor taxes to finance the subsidy. Thus, in a second-best setting, an abatement subsidy suffers an efficiency disadvantage relative to an emissions tax that exploits the revenue-recycling effect because the subsidy must be financed through costly distortionary taxes[49].

In a model where firms' production technologies do not exhibit constant returns to scale, an abatement subsidy can have an additional efficiency disadvantage by inducing excessive entry (i.e., too many firms)[50]. To avoid such additional efficiency costs, the baseline e_0 on which the subsidy is calculated should be positive only for firms that, in the baseline, are actually generating emissions. Potential new entrants should have a value of zero for e_0. However, political considerations might tempt regulators to allow new entrants to enjoy the subsidy, which would require assigning positive values of e_0

to a setting involving government regulation and pre-existing distortionary taxes, it becomes important to examine how different property-rights specifications are related to the acquisition and disposition of government revenue. We shall show that the polluter-pays principle has a potential efficiency advantage once these revenue issues are taken into account.

[49] Parry (1998) was the first to provide an analytical general-equilibrium treatment of the efficiency cost of abatement subsidies in the presence of distortionary factor (labor) taxes. His analysis of a subsidy to reduction in production of a "dirty" consumption good yields results similar to those we describe here. Parry considers other types of subsidies as well, including a subsidy to a non-polluting ("clean") consumption good. As shown by Fullerton (1997), this latter subsidy policy is functionally identical to (and produces the same efficiency impacts as) a tax on the dirty consumption good.

[50] The present model implicitly assumes that production involves constant returns-to-scale and that the abatement subsidies are awarded only to existing firms.

to such entrants. As pointed out by Baumol and Oates (1988) and Pezzey (1992), doing so leads to excess entry. Under these circumstances, the lump-sum component of the subsidy, te_0, is no longer truly lump sum, because e_0 now depends on firms' decisions whether to enter the market.

Beyond these differences in efficiency, there are important differences between abatement subsidies and emissions taxes in terms of distribution, where that difference is represented by the lump-sum transfer sKe_0. In their roles as taxpayers, individuals would abhor the subsidies; but as owners of polluting enterprises, they would embrace them.

4.1.4. Performance standards

Much environmental regulation takes the form of performance standards – ceilings imposed on the amount of pollution emissions per unit of output. Examples include automobile tailpipe-emissions requirements and water-quality regulations that impose ceilings on effluent-output ratios. The performance standard can be represented by the constraint $E/ND \leqslant \bar{\varepsilon}$, which we assume is binding.

Firms maximize profits subject to the performance constraint. The Lagrangian function associated with this profit-maximization problem is

$$d(p_D - w_p) - aw_p - \lambda^e(e/d - \varepsilon), \tag{58}$$

where $d \equiv ND/K$ = per-firm production of the dirty good, $a \equiv A/K$ = per-firm emissions abatement, $e \equiv E/K$ = per-firm emissions (as previously), p_D is the price of the dirty consumption good, and w_p is again the producer wage (which, as before, is normalized to 1). To gauge the efficiency impacts of a performance standard relative to an emissions tax, it is useful to establish that this instrument is equivalent to the revenue-neutral combination of an emissions tax and a subsidy to production of the dirty good D. To see this, notice that under the combination of emissions tax t_E^R and production subsidy s^R, the firm's profit function is

$$d(p_D - w_p) - aw_p - t_E^R e + s^R d. \tag{59}$$

Revenue-neutrality requires that $s^R ND = t_E^R E$ or, equivalently, $s^R = t_E^R \varepsilon$. With $t_E^R = \lambda^e$, the firm's maximization problem under the tax/subsidy policy becomes identical to the Lagrangian under the performance standard, which establishes the equivalence between the two policies. Thus, the subsidy component of the performance standard constitutes the difference between a performance standard and a pure emissions tax. This component gives rise to an additional efficiency cost relative to that of the emissions tax. The added cost arises because the subsidy component makes the price of the dirty consumption good too low from an efficiency point of view[51].

[51] Goulder et al. (1999) show that at the initial incremental unit of abatement, the performance standard has no efficiency disadvantage relative to the emissions tax. This is the case because the source of the disadvantage – namely, the subsidy component s^R – approaches zero as the level of abatement and t_E^R approach zero. For a related discussion, see Fullerton and Metcalf (2001).

As will be discussed below, it is possible to rectify this problem by combining a performance standard with a tax on the dirty consumption good.

4.2. *Uncertainty and instrument choice*

Much of the preceding discussion suggests that environmental taxes enjoy significant efficiency advantages over other instruments for environmental protection. In the real world, however, environmental taxes are employed much less frequently than are technology-based regulations or performance standards. This may partly reflect the inability of lawmakers and the general public to appreciate the efficiency virtues of environmental taxes. It may also attest to the tendency of the political process to avoid the distributional impacts that would stem from emissions taxes, even when the efficiency virtues of such taxes are acknowledged. Still, the reluctance to embrace environmental taxes could reflect some efficiency *dis*advantages of emissions taxes that can arise in more complex settings than those considered so far. The presence of uncertainty, in particular, adds further dimensions to instrument choice, and in some circumstances may militate in favor of non-tax approaches. Moreover, uncertainty and associated costs of monitoring and enforcement may make taxes on output preferable to taxes on emissions, despite the fact that emissions taxes are more closely connected to the externality in question. We take up these issues in this subsection.

4.2.1. *Instrument choice under imperfect or costly monitoring*

4.2.1.1. Imperfect monitoring and the choice between emissions taxes and emissions quotas. In general, regulators lack perfect information as to the extent to which particular firms are complying with pollution-abatement rules. Under such circumstances, firms may exceed applicable pollution standards or they may under-report emissions in submitting emissions-tax payments. Harford (1978) analyzed the behavior of risk-neutral firms under pollution quotas (standards) or taxes in this setting[52]. In Harford's model, the government imposes fines on firms that are found to violate pollution regulations. The expected penalty (the product of the probability of detection and the size of the fine if a violation is detected) is an increasing function of the level of violation. Under pollution quotas, firms choose a level of emissions that equates the marginal increase in the expected fine with the marginal benefit (cost reduction) from a higher level of pollution. If the marginal penalty is an increasing function of the size of the violation, a tighter pollution standard raises the marginal penalty associated with any given level of pollution and therefore implies that firms will optimize at a lower level of pollution.

Under pollution taxes, the firm chooses both the actual level of pollution and the reported amount of pollution. Tax payments are based on reported pollution. The

[52] Other studies of environmental regulation under imperfect or costly monitoring include Downing and Watson (1974), Harrington (1988), Lewis (1996) and Swierzbinski (1994).

violation is the difference between actual and reported pollution. Harford finds that it is optimal for firms to choose levels of actual pollution at which marginal abatement costs equal the tax rate[53]. The scale of the penalty function affects reported pollution, but not actual pollution. This suggests a potential advantage of emissions taxes over quotas when emissions cannot be perfectly monitored. If the emissions tax rate is set optimally, then the tax will generate the efficient level of actual pollution. In contrast, there is no simple way to induce an efficient pollution level under the quota.

4.2.1.2. Costly monitoring and the choice between emissions taxes and output taxes. Imperfect or costly monitoring can also affect the choice between emissions taxes and output taxes. An attraction of emissions taxes is that they produce both input-substitution and output-demand effects that contribute to efficient emissions reductions. The input-substitution effect is the substitution of non-polluting inputs for inputs associated with pollution. The output-demand effect is the substitution of other goods or outputs for the (now higher-priced) good whose production involves pollution. While emissions taxes generate both effects, output taxes produce only the output-demand effect. In the absence of uncertainty, this makes output taxes less efficient than emissions taxes as instruments for reducing emissions.

As pointed out by Baumol and Oates (1988), output taxes may have a compensating advantage because it may be less costly to monitor output than to monitor emissions. This potential advantage must be weighed against the disadvantage of omitting the input-substitution effect. Schmutzler and Goulder (1997) employ a model in which monitoring emissions is costly but monitoring output involves no cost. They find that, depending on the scope of monitoring costs, the input-substitution effect, and the output-demand effect, the optimal policy will involve either pure emissions taxes, pure output taxes, or a mix of the two. Higher (lower) costs of monitoring emissions, a smaller (larger) input-substitution effect, and a larger (smaller) output-demand effect contribute toward the optimality of pure output (emissions) taxes. If marginal monitoring costs are not too high, a mix of output and emissions taxes may be optimal. This can be seen if one considers starting with a pure emissions tax and then reducing slightly the emissions tax rate while incrementing the output tax (from zero) in a way that keeps emissions constant. The initial substitution of the output tax for the emissions tax does not change the cost (ignoring monitoring costs) of achieving emissions reductions because the significance of losing the input-substitution effect is initially zero. At the same time, substituting the output tax for part of the emissions tax

[53] Specifically, it is optimal for the firm to choose a level of emissions such that marginal abatement costs (or the marginal benefit from emissions) equal the marginal expected penalty. It is also optimal to equate the marginal expected penalty with the tax rate. By transitivity, marginal abatement costs should equal the tax rate at the optimum.

yields a first-order saving in monitoring costs. Hence the mixed policy is superior to the pure emissions tax[54].

4.2.1.3. Using two-part instruments to overcome monitoring problems. Monitoring costs can be reduced, and efficiency enhanced, by employing two-part instruments. In Subsection 4.1.4 we observed that a performance standard is equivalent to an emissions tax plus a subsidy to output. This implies that a performance standard, combined with an appropriately scaled tax on output, is equivalent to an emissions tax. The output-tax component of this two-part instrument neutralizes the efficiency disadvantage of the performance standard relative to the emissions tax.

Eskeland and Devarajan (1995) analyze an option like this in the context of air pollution in Mexico City. They demonstrate that adding a tax on gasoline to a system involving mandated automobile pollution-reduction technologies yields efficiency gains, and that the resulting two-part system approximates the impact of a tax on automobile emissions[55]. The addition of the gasoline tax is equivalent to the removal of the mandated technology's implicit subsidy to automobile use; hence it helps remove the efficiency disadvantage of the mandated technology relative to a tax on emissions.

A deposit-refund system is another important example of a two-part instrument designed to overcome monitoring problems. Under such a system, consumers pay a surcharge when purchasing products whose improper disposal would lead to environmental harm. The deposit is refunded if the consumer returns the product to an approved center for recycling or proper disposal. Thus the refund component helps overcome the difficulty of monitoring improper disposal[56].

Fullerton and Wolverton (1997) show that difficulties of monitoring "dirty" inputs can be overcome through the combination of a tax on output and subsidy to "clean" inputs. This two-part instrument is equivalent to a tax on the "dirty" input. Thus, for

[54] Introducing an output tax instead of an emissions tax is just one example of how regulators can tax activities related to, but imperfectly associated with, emissions. A general examination of this issue is provided by Wijkander (1985), who emphasizes that efficiency depends not only on how taxed (and subsidized) activities are related to emissions, but also on how they are related to each other. Cross-effects between related goods can yield counterintuitive results. In particular, a subsidy (rather than a tax) on a complement to emissions may be optimal. The reason is that it alleviates the distortions due to the imperfect link between emissions and another complementary good, which is taxed.

[55] Some of the differences between their two-part instrument and an emissions tax may be due to the fact that they combine an output tax with a mandated technology rather than a performance standard. The mandated technology does not provide the same incentives for input substitution inherent in a performance standard. For further analysis of this issue see Goulder, Parry, Williams and Burtraw (1999).

[56] For a theoretical exposition of deposit–refund systems, see Bohm (1981). Such systems have been implemented in the USA, Canada, and some European countries through "bottle bills" intended to control litter from beverage containers and reduce the flow of solid waste from landfills. See Menell (1990). In the USA a deposit-refund system has been applied to lead-acid batteries as well.

example, if the use of coal by electric power plants is difficult to monitor, the effect of a coal tax can be duplicated by the combination of a tax on electricity output and a subsidy to all inputs to electricity other than coal. Fullerton and Wolverton point out that a deposit-refund system is much like this combined tax and subsidy. The deposit on batteries is like an output tax, and the refund for proper battery disposal is akin to a subsidy to a clean "input" (a clean method for using and disposing of the battery).

4.2.1.4. Liability rules as alternatives to taxes in the presence of uncertainty. One can think of environmental damages as a function of various activities by households or firms. Consider in particular the case where the damage or harm, h, is a function of the vector $\boldsymbol{x}$ of actions taken by a firm: $h = f(\boldsymbol{x})$. The vector $\boldsymbol{x}$ might represent productive inputs; it could also represent a vector of emissions or other by-products from production. The important distinction is between various firm-level phenomena (given by $\boldsymbol{x}$) and the harm h that results from them. If the function f is known by the regulator, then in principle the regulator could achieve the same outcome either by taxing the harm h or by taxing different vectors $\boldsymbol{x}$ according to the harm that they generate.

Typically, regulators will not know the function f with certainty. Moreover, the harm h associated with a given vector $\boldsymbol{x}$ might have a stochastic component. In this case, one can associate with each $\boldsymbol{x}$ a distribution of harms, H, where $H = g(\boldsymbol{x})$.

In this situation the regulator could discourage harm through a tax on $\boldsymbol{x}$ (for example, a tax on fuel inputs or on emissions). The regulator will succeed in bringing about efficient choices of $\boldsymbol{x}$ if the tax equals the expected harm associated with $\boldsymbol{x}$: that is, if $t(\boldsymbol{x}) = E(g(\boldsymbol{x}))$ for all $\boldsymbol{x}$ [57], and economic agents (firms and potential victims) are risk-neutral.

However, the regulator may have very little information as to the distribution of harms associated with each $\boldsymbol{x}$. An alternative is to implement liability rules: penalties based on the actual harm, h [58]. A potential advantage of such rules is that they do not require the regulator to observe $\boldsymbol{x}$ or know the function g. The regulator only needs to be able to observe the harm h when it occurs and trace it to the responsible firm. In contrast, to achieve efficient regulation through a tax on $\boldsymbol{x}$, the regulator must be able to monitor $\boldsymbol{x}$ perfectly and must know the function $E(g(\boldsymbol{x}))$.

One can imagine circumstances when liability rules will be considerably more attractive to the regulator. For example, when $\boldsymbol{x}$ is a large vector – as when potential harm derives from activities of the firm along a great many dimensions (safety, personnel, upkeep, etc.) – it might be especially difficult to devise a tax that captures

[57] We abstract from the issue of prior distortionary taxes. In the presence of such taxes, efficiency is achieved when $t(\boldsymbol{x}) = E(g(\boldsymbol{x}))/\text{MCPF}$.

[58] For a further discussion of the economics of liability rules, see Section 2 of chapter 25 by Kaplow and Shavell in this volume and the references cited therein. This discussion concerns what are termed *strict* liability rules. These contrast with negligence rules and other legal provisions assigning liabilities based on harm. See Kaplow and Shavell (2001) for a discussion.

expected harm. This is the case both because the regulator would not be able to observe $\boldsymbol{x}$ and because it would have a very inexact idea of how this complicated behavior affects the probability of harm. Circumstances of this sort are fairly common, which may help explain why liability rules sometimes can be more important in dealing with externalities (in environmental contexts and elsewhere) than environmental taxes. At the same time, in some circumstances liability rules may be less attractive than taxes. Such circumstances seem to apply in the paradigmatic case of pollution-generation – where the source of harm is an identifiable pollutant stemming from a production plant. In this case, it may be easier to track $\boldsymbol{x}$ (which may be a scalar) than the produced harm (which may be difficult to trace back to its source). In this case a pollution tax has an advantage over a liability rule.

4.2.2. *Uncertainty about costs and benefits and the choice between price-based and quantity-based instruments*

We have analyzed how, in a second-best setting, the ability of emissions taxes to raise revenue and exploit the revenue-recycling effect yields an important efficiency advantage of such taxes over non-auctioned emissions quotas. In the presence of uncertainty about abatement costs or environmental damages, further issues arise that can either weaken or strengthen the case for emissions taxes relative to emissions quotas.

These issues were first addressed formally by Weitzman (1974)[59], who considered the setting where regulators are uncertain as to the marginal costs and marginal benefits of pollution abatement, and must choose between a price-based instrument (that is, a pollution tax) and a quantity-based instrument (that is, a pollution quota or an imposed level of pollution abatement). Weitzman's basic results are heuristically presented in Figures 6a,b. Figure 6a displays the case where regulators are uncertain as to the costs of emissions abatement. In the diagram, MC_E and MC_R respectively stand for the expected and actual (or realized) marginal costs of abatement, and MB represents the known marginal benefits (avoided damages) from abatement. Regulators must either set an emissions tax t or require a given quantity of abatement, a. (Setting the quantity a is the same as specifying the emissions quota $\bar{e}$, where $\bar{e} = e_0 - a$, and e_0 represents emissions in the absence of regulation.) Regulators are regarded as risk-neutral and aim to maximize expected net benefits from emissions reductions.

If the uncertainty as to the position of the MC schedule is symmetric, the tax rate that is optimal *ex ante* is the rate t^* in the diagram, and the quantity of abatement that is optimal *ex ante* is a^*. These levels equate marginal benefits with the expected marginal costs. *Ex post*, however, neither the tax nor the restriction on the quantity of

[59] Weitzman's central results were anticipated in earlier papers by Lerner (1971) and Upton (1971), who employed somewhat less formal analyses. Additional analyses include Adar and Griffin (1976), Fishelson (1976), Roberts and Spence (1976) and Stavins (1996). Newell and Pizer (1998) extend this uncertainty framework to a dynamic setting that considers connections between pollution emissions (flows) and concentrations (stocks).

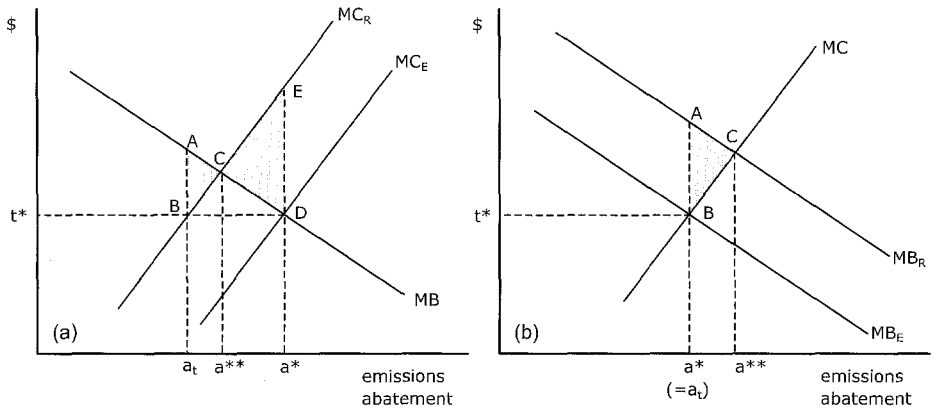


Fig. 6. Uncertainty and the choice of policy instrument: (a) cost uncertainty; (b) benefit uncertainty.

abatement (or emissions quota) is optimal, since neither promotes abatement at a level that equates *realized* marginal costs (MC_R) with marginal benefits. With the realized marginal cost schedule MC_R in Figure 6a, the optimal level of abatement is a^{**}, which differs from a^* and from a_t, the level of abatement that results under the tax t^*. (When the tax is t^*, a_t is optimal because it equates firms' actual marginal abatement costs to this tax rate.) In Figure 6a, the efficiency losses relative to the *ex post* optimum are shown by the shaded triangles. The tax implies the loss represented by the triangle ABC; the quantity regulation implies the loss given by the triangle CDE. In this case, the loss is substantially larger under quantity-based regulation than under the tax. This reflects the relative steepness of the MC curve in comparison with the MB curve. When the MB curve is relatively flat, the tax avoids especially large (and costly) errors in the quantity dimension. It is easy to show diagrammatically that if the MC curve is relatively flat in comparison with the MB curve, the quota leads to smaller expected losses (relative to the *ex post* optimum) than the tax. These results are confirmed and generalized in Weitzman's mathematical analysis. The case of a relatively steep MB curve applies in the neighborhood of serious threshold effects. If global climate change, for example, is characterized by a threshold where small increases in emissions would cause significant climate change, then near this threshold the marginal-benefit curve will be very steep. In such a situation, quotas or other quantity-based instruments may be preferred.

In contrast, uncertainty on the benefit side does not discriminate between taxes and quotas (quantity-based regulation) if this is the only uncertainty or if benefit uncertainty is uncorrelated with the cost-side uncertainty. This is illustrated in Figure 6b, which assumes no uncertainty on the cost side. Here the efficiency loss is ABC under both policies. When there is no uncertainty as to marginal costs, the abatement level a_t chosen by firms under the tax is the same as a^*. Hence the choice of instrument has no bearing on the efficiency loss. If the uncertainty and cost uncertainty are correlated,

however, uncertainty on the benefit side affects the relative attractiveness of quotas and taxes[60].

This analysis has ignored the second-best issues discussed earlier. Second-best considerations have no bearing on the choice, under uncertainty, between emissions taxes and *auctioned* quotas, since auctioned quotas and emissions taxes are equivalent in terms of their tax-interaction and revenue-recycling effects. Thus, Weitzman's analysis provides a useful guide to the choice between taxes and auctioned quotas in both first- and second-best settings[61]. On the other hand, second-best considerations give a premium to emissions taxes over *grandfathered* quotas, for reasons discussed earlier. Uncertainty considerations of the sort we have just discussed could either reinforce or offset this premium.

It is important to recognize that the Weitzman analysis assumes that the government must impose a linear tax on emissions. The choice between taxes and quotas is different when the government can make use of nonlinear taxes. As shown by Roberts and Spence (1976) and Kaplow and Shavell (1997), a nonlinear tax on emissions dominates an emissions quota in the presence of uncertainty about abatement costs and emissions damages. *Ex ante,* it is optimal for the government to introduce a tax schedule which duplicates the schedule of expected marginal damages as a function of emissions. The nonlinear tax schedule expresses a more complex relationship between emissions and damages than can be expressed under either a linear tax or a quota, and this helps it reduce the errors associated with the uncertainty about the actual position of marginal cost curve. Although nonlinear tax schemes are seldom introduced in practice, Kaplow and Shavell contend that such schemes need not be difficult to administer.

5. Distributional considerations

5.1. Efficiency–equity trade-offs

By assuming homogeneous households, the previous sections abstracted from distributional considerations. Policies were analyzed mainly in terms of efficiency. Now we consider the case where the government is concerned with distributional impacts as well as efficiency.

In environmental policy making, a trade-off often emerges between efficiency and distribution (equity). To illustrate this, we expand the model from Section 2 to include two types of households. The first type, which will be called the "active" household,

[60] This point was briefly noted by Weitzman. Stavins (1996) explores this issue in detail.

[61] Second-best considerations could influence the choice between emissions taxes if they affected the slopes of the relevant marginal cost or marginal damage curves. However, in a partial-equilibrium framework Schöb (1996) shows that pre-existing taxes do not alter the relevant slopes, so that Weitzman's first-best choice rule between price and quantity regulation remains valid in a second-best setting.

relies entirely on labor income. The second type, the "inactive" household, obtains income only from government transfers. We assume that transfers are not subject to the tax on labor income but are subject to taxes on consumption[62]. Under these circumstances, higher taxes on consumption reduce the purchasing power of "inactive" households. The relative changes in the demands for the two commodities (45) and (46) become

$$\tilde{C} = (1-\alpha_Y)(\tilde{w}_R + \tilde{L}) - \alpha_Y(1-\alpha_C)\tilde{t}_D + (1-\alpha_C)\sigma_J\tilde{t}_D, \tag{60}$$

$$\tilde{D} = (1-\alpha_Y)(\tilde{w}_R + \tilde{L}) - \alpha_Y(1-\alpha_C)\tilde{t}_D - \alpha_C\sigma_J\tilde{t}_D, \tag{61}$$

where α_Y denotes the share of non-labor income in aggregate household income (after labor taxes). Abstracting from pollution taxes on intermediate inputs, we arrive at the following expression for employment:

$$\tilde{L} = \frac{\varepsilon_{LL}^U}{\Delta_T}\left[-\theta_D\alpha_C(1-\alpha_C)\,\omega_L(1-t_L)\sigma_J + \alpha_Y S(1-\alpha_C)\right]\tilde{t}_D, \tag{62}$$

where

$$\begin{aligned}\Delta_T \equiv\ & (1-\alpha_Y)(1-\theta_L)\,\omega_L\left[1-\theta_D(1-\alpha_C)\right] \\ & -\varepsilon_{LL}^U\left[\theta_L\omega_L(1-\alpha_Y) + \omega_L\alpha_Y + \theta_D(1-\alpha_C)\,\omega_L(1-\theta_L)(1-\alpha_Y)\right].\end{aligned} \tag{63}$$

Expression (62) indicates that in the absence of government transfers, an incremental environmental tax reform does not affect employment if the initial pollution tax is zero. This corresponds to the earlier results from Section 3. However, if transfers are positive such a reform boosts employment [see Equation (62) with $\theta_D = 0$]. In this case the government is able to more than compensate workers for the real income loss due to a higher environmental tax because the tax reform ends up shifting the tax burden from workers to transfer recipients. Thus, real wages and the labor supply increase. Accordingly, the increase in environmental quality is accompanied by a higher level of employment, thereby improving efficiency and reducing the labor-market distortion due to the labor tax[63].

Bovenberg and de Mooij (1994b) offer further analysis of this issue and consider in particular a reform that begins with positive initial taxes (i.e. $\theta_D > 0$). They show that the environmental tax reform can increase employment (and produce the double dividend) only if the reform involves a reduction in the real value of transfers and thus redistributes purchasing power from transfer recipients to wage earners. Their analysis points out a rather robust trade-off between efficiency and equity[64].

[62] Transfer recipients are not compensated for consumption taxes because the price index used to determine real transfers does not include consumption taxes.

[63] Here an efficiency improvement is identified with a potential Pareto improvement.

[64] This model indicates, more generally, that efficiency gains can be reaped only by reducing the real value of transfers. The government need not employ the environmental tax reform to achieve such gains. In this model, the government could reap these efficiency gains more directly by cutting nominal transfer payments, by subjecting transfers to the labor-income tax, or by replacing a tax on labor income by a broad-based tax on consumption.

The analysis above abstracted from the distribution of environmental benefits. Environmental taxes differ from other taxes because they not only finance ordinary public goods but also augment the supply of the environmental public good. The distribution of environmental benefits can be such as to create scope for an efficiency-enhancing reform, even when the government must meet the constraints of revenue- and distributional neutrality. For example, if environmental benefits accrue especially to the "inactive" household, these benefits can offset the welfare impact associated with a reduction in the real value of transfers. Hence, under these circumstances, the government might be able to introduce a policy that reduces the real value of transfers while still satisfying the distributional constraint that the overall welfare of this household be maintained. This example illustrates how the distribution of environmental benefits from pollution taxes can potentially "grease the wheels" of tax reform.

5.2. The Pigouvian rule reconsidered

The Pigouvian rule – to set taxes equal to marginal environmental damages – applies the Samuelson condition to the public good of the environment. As noted, if governments have access to individual-specific lump-sum taxes, the Samuelson condition holds: it is optimal for the marginal rate of transformation between the public and private goods to be equal to the sum of the marginal rates of substitution.

5.2.1. Conditions that would resurrect the Pigouvian rule

In practice, governments do not have access to individual-specific lump-sum taxes because they cannot observe individual abilities. Instead, they have to rely on observable behavior (i.e., labor income) to distinguish between various households[65]. The previous analysis suggested that in the absence of lump-sum taxes, the Samuelson condition (Pigouvian rule) is no longer optimal. However, in an expanded analysis that considers distributional concerns, the Samuelson condition may apply after all in some special circumstances [see Christiansen (1981), Boadway and Keen (1993), Kaplow (1996), Pirttilä and Tuomala (1997) and Slemrod and Yitzhaki (2000)]. The literature on public-goods provision has derived the following three conditions which together assure that the Samuelson rule continues to hold even if governments cannot employ individual lump-sum taxes:

- The government can impose a nonlinear income tax system whose rates can be adjusted to offset the distributional effects of the provision of additional public goods.
- Households have identical tastes.

[65] Specifically, the government observes only labor income and cannot separately observe the wage rate or hours worked.

– In homogeneous utility, leisure is weakly separable from (public and private) goods.

When these conditions hold, the optimal level of provision of the public good is given by the Samuelson rule. The reason is as follows. Suppose the government applies the Samuelson condition to determine the level of provision of a public good, and that it provides this good in a way that is distributionally neutral; that is, the (nonlinear) tax system is adjusted so that the benefit enjoyed by any individual from higher environmental quality is exactly offset by the additional taxes paid. If the three conditions hold, then financing the public good in this way leaves each person's incentives unchanged on the labor–leisure margin, and thus introduces no additional distortion. Thus the Samuelson rule (which, in the context of environmental public goods, is the Pigouvian rule) is optimal. The separability condition is necessary to ensure that the ratio between private and public goods does not alter the marginal utility of leisure[66].

Kaplow (1996) points out the optimality of the Pigouvian rule in the special case where the third condition above is satisfied because environmental quality and private goods are separable from leisure in utility. This case is a natural benchmark because it provides the conditions under which efficiency and equity can be separated by perfectly matching targets and instruments. In this case, the income tax takes care of distributional concerns, which allows pollution taxes to be aimed solely at internalizing externalities. Under the conditions specified by Kaplow, the impacts on leisure demand from the provision of the public good and from the taxes that finance it exactly offset one another. Hence financing the public good has no distortionary impact, and the "first-best" Pigouvian rule applies.

The Samuelson condition also continues to apply if public goods enter production as a separable intermediate input (i.e., the marginal products of other inputs are unaffected by the level of public goods) and if consumption taxes can be set set optimally [see Christiansen (1981)]. This result is closely related to the result derived by Diamond and Mirrlees (1971) that distributional concerns do not justify violating production efficiency if the government can optimally adjust consumption taxes. Intuitively, distributional issues are more efficiently addressed with consumption taxes, which directly affect consumer prices, than with taxes on intermediate goods, which influence consumer prices only indirectly.

5.2.2. Difficulties in meeting those conditions

In practice, however, the conditions that would restore the Pigouvian rule are difficult, if not impossible, to obtain. The following issues seem especially relevant.

[66] This is closely related to the familiar result derived by Atkinson and Stiglitz (1976) that commodity taxes should be uniform if produced goods are weakly separable from leisure in utility. Intuitively, this latter condition ensures that, compared to the income tax, commodity taxes are less efficient instruments for redistribution. Accordingly, these taxes can be targeted at efficiency, implying uniform commodity taxes.

5.2.2.1. Imperfect compensation. Because of information problems and associated administrative costs, the type of nonlinear tax system described above may be infeasible. Governments may find it difficult to identify the workers that suffer inordinately from environmental taxation and to calculate the required adjustment of the income tax to leave the distribution unaffected. The nonlinear tax system prescribed above would thus face serious information problems. These systems may be difficult to administer as well, in part because of the complexities involved in adjusting the tax schedule in the face of new information[67]. Moreover, a nonlinear income tax will not be able to compensate all agents for the effects of the pollution tax if environmental preferences vary within income classes. Finally, governments may not be able to adjust income taxes to offset the distributional effects of environmental policy because of political and institutional constraints.

The absence of sufficient instruments to compensate distributional effects implies that pollution taxes cannot be targeted solely at internalizing pollution. Indeed, the inability to offset the distributional effects of pollution taxes is one of the main obstacles to the introduction of pollution taxes[68].

5.2.2.2. Non-separability of utility. If the environmental public good and private goods are not separable from leisure, the Pigouvian rule is no longer optimal. Under such circumstances, additional provision of the environmental public good affects the marginal rate of subsitution between leisure and consumption and thereby influences labor supply[69]. If, in particular, environmental quality is a weaker complement to leisure than private goods are, an improvement in environmental quality makes work relatively more attractive, yielding an increase in labor supply and a reduction in the labor-market distortion. Hence, under these circumstances the environmental public good should be provided at a higher level than that endorsed by the Pigouvian rule[70].

[67] In the absence of nonlinear income taxes, commodity taxation must play a redistributive role. Deaton (1977) investigates how in this case the optimal commodity tax structure should strike a balance between equity and efficiency considerations.

[68] As pointed out by Feldstein (1976), the distributional issues associated with a reform of an existing tax system are rather distinct from those associated with designing a tax system from scratch.

[69] Cremer et al. (1998) explore how externalities affect the optimal structure of the nonlinear income tax. They show that externalities may change the formula for the optimal marginal income tax rates if commodity transactions are anonymous and the government therefore cannot levy nonlinear commodity taxes. Intuitively, in the absence of sufficiently rich instruments to control the externality directly through a tax on pollution, it is second-best optimal to address the externality indirectly through the nonlinear income tax.

[70] This result is consistent with expression (25). According to this expression, the lower tax should be imposed on the commodity that is less complementary to leisure. Hence, this commodity should be 'overprovided' in the sense that its marginal rate of transformation should exceed its marginal rate of substitution.

To consider this more closely, note that if environmental quality is less complementary to leisure than private consumption goods are, the marginal willingness to pay for the environment (i.e. the marginal rate of substitution between environmental quality and private goods) declines with the amount of leisure. This relaxes the self-selection constraint that restricts the amount of redistribution [see Boadway and Keen (1993)]. In particular, if high-ability households mimic the low-ability households by collecting the same income, the only difference between the households is that the high-ability households enjoy more leisure. Their marginal willingness to pay for the environment is lower than that of low-ability households because this willingness to pay declines with leisure. Accordingly, raising the ratio of environmental quality to private commodities raises utility of low-ability households compared to that of the mimicking high-ability households. By relaxing the self-selection constraint, a higher environmental tax allows for a less progressive tax system – that is, a lower (distortionary) marginal labor tax. In this way, distributional concerns promote a higher level of provision of the environmental public good[71]. However, these concerns negatively affect the provision of environmental quality if, compared to private consumption of produced commodities, environmental quality is more complementary to leisure.

5.2.2.3. The environment as an intermediate input. The Pigouvian rule also is suboptimal if environmental quality and productive inputs are not separable. Under these circumstances, the provision of the environmental public good has distributional implications through effects on the relative wages of workers of different skills. If, in particular, environmental quality is more complementary to labor provided by low-ability households ("unskilled" labor) than to labor provided by high-ability households ("skilled" labor), enhancing environmental quality redistributes well-being in favor of unskilled labor. To the extent that society values this redistribution, it would support a higher level of environmental quality than that prescribed by the Pigouvian rule. This circumstance could conceivably apply to policies that improve water quality in marine fisheries. This improves the productivity of fishermen, and the resulting distributional impact could support a higher level of regulation than otherwise would be considered justified[72].

5.3. Re-examining instrument choice in light of distributional issues

If the three sufficient conditions for the optimality of the Samuelson rule hold (see Subsection 5.2.1), efficiency and distributional concerns are best addressed by separate

[71] A similar argument holds if preferences are not homogeneous but instead vary with unobservable ability [see Mirrlees (1976)]. In particular, the environment should be "overprovided" if low-ability households feature relatively strong preferences for the environment.

[72] For further analysis of this issue, see Boadway and Marchand (1995).

policy instruments. The choice of environmental policy instrument, in particular, should be based only on efficiency considerations. Under these circumstances there is a clear combination of instruments that best meets efficiency and distributional objectives. The optimum here involves providing the level of environmental quality dictated by the Pigouvian rule, supplying other public goods according to the Samuelson condition, and employing a nonlinear income tax system that deals with distributional concerns[73].

In general, the three conditions are unlikely to hold, however. Yet the analyses that identified these conditions provide important lessons relevant to other circumstances. They reveal the importance of considering factor-supply effects stemming from the changes in the level of the environmental public good. In addition, they indicate that, where possible, it is useful to assess efficiency impacts subject to the constraint of "distributional neutrality" – the requirement that the distribution of individual well-being remain unchanged across the policies under consideration. Moreover, they suggest that when the policies under consideration are not distributionally neutral – a typical situation when actual policy alternatives are involved – differences in distributional effects are relevant to the choice of environmental policy instrument. The relative attractiveness of a given policy instrument will depend not only on efficiency effects but also on distributional impacts and the weights given to those impacts in the social welfare function[74].

Much will depend on whether the government has sufficient instruments to meet both distributional and efficiency objectives. Environmental taxes, in particular, may become more attractive to the extent that they are part of a larger tax-reform package. When it combines new environmental taxes with other reforms that address distributional issues, the government utilizes a large number of policy instruments, and the potential for Pareto-improving outcomes increases.

Potential trade-offs between efficiency and distribution (equity) become relevant to the choice among policy instruments identified in Section 4. We observed that in a second-best setting, revenue-raising instruments such as environmental taxes have a potential efficiency advantage over policies like freely offered tradeable permits that do not raise revenue. The former policies exploit this advantage to the extent that revenues are used to finance reductions in the rates of pre-existing distortionary taxes.

[73] When the three sufficient conditions hold, the optimal level of environmental quality can be provided by emissions taxes or emissions quotas. Adjustments to the nonlinear income tax would offset what otherwise would be an efficiency disadvantage of quotas, namely, their inability to exploit the revenue-recycling effect.

[74] If, in particular, the government previously achieved an optimal distribution of income, then an incremental lump-sum redistribution to any individual has a value equal to the MCPF. Under these conditions, the quota policy's rents also have an incremental value equal to the MCPF. Hence the efficiency disadvantage of the quota would be exactly offset by the value of its distributional impact. See Kaplow (1996) for discussion of related issues.

Buchanan and Tullock (1975) showed that environmental policies, by causing restrictions in output, have the potential to generate significant rents to the regulated firms. The potential to enjoy significant rents has distributional significance and thus is relevant to the choice among environmental policy instruments. Consider, for example, two policies involving tradeable emissions permits. Under one policy, all of the permits are auctioned out to industrial sources of pollution; under the other, the permits are freely provided to these firms. The latter policy enables firms to retain as private rents what otherwise would be government revenue. Buchanan and Tullock pointed out that the latter policies can cause firms' profits to be higher than they would be in the absence of regulation. In keeping with this observation, Bovenberg and Goulder (2001) found that very modest grandfathering (free provision) of emissions permits is consistent with preserving profits. Under a policy in which tradeable permits are employed to limit US emissions of carbon dioxide, only a small percentage (around 10 percent) of the permits must be given out free in order to preserve profits of the regulated fossil-fuel industries; a very large percentage can be auctioned. When most of the permits are auctioned, the government's sacrifice of revenue is small and thus the sacrifice of efficiency (relative to the case of 100% auctioning – the most cost-effective case) is small as well. Only a small share of the permits must be freely provided because the policy produces large potential rents. Firms only need to retain a small share of the potential rents to maintain profits.

Thus, policy makers can address distributional and efficiency objectives by deciding what fraction of potential revenues from an environmental policy will actually be collected. More generally, by introducing more complex policies (including policies in which more than one instrument is invoked), the government gains flexibility in attending to both efficiency and distributional concerns. This may help policies come closer to achieving Pareto improvements and thereby enhance political feasibility.

6. Summary and conclusions

This chapter has analyzed economic issues surrounding the use of taxes and other instruments for environmental protection. It attests to the importance of general-equilibrium effects – in particular, interactions between environmental policy initiatives and pre-existing distortionary taxes. Because of these interactions, some of the most important efficiency impacts of environmental policies take place outside of the sector, industry, or market that is targeted by the regulation.

Two central ideas explain these interactions and form the basis of many of the results discussed in this chapter. First, environmental taxes and other forms of environmental regulation act as implicit taxes on factors of production because they raise the costs and prices of produced goods relative to the prices of factors, thereby lowering real factor returns. Second, insofar as they function as implicit taxes on factors, environmental taxes and regulations compound distortions posed by pre-existing factor taxes.

6.1. Optimal tax issues

These ideas underlie, for example, the main results on the optimal setting of environmental taxes. In a second-best setting where distortionary taxes represent a necessary source of revenue, the optimal rate for an environmental tax typically is less than the Pigouvian rate. This reflects the fact that environmental taxes compound the distortions of pre-existing taxes – even after accounting for the value of the revenues raised by these taxes. Consequently, a given environmental tax rate entails a higher cost than it would in a first-best world. Hence the optimal tax rate is lower.

6.2. Costs of revenue-neutral reforms

Tax interactions also explain why revenue-neutral environmental tax reforms (as opposed to optimal tax policies) tend to be more costly in a second-best setting than in a first-best world. In examining revenue-neutral reforms, it was useful to decompose the overall impact on gross cost into a *tax-interaction* effect and a *revenue-recycling* effect. Typical revenue-neutral environmental tax reforms produce a negative (in efficiency terms) tax-interaction effect by raising overall output prices and thereby lowering returns to factors. Such reforms also produce a positive revenue-recycling effect to the extent that they finance reductions in marginal tax rates of pre-existing distortionary taxes. A main lesson from analyses of revenue-neutral environmental tax reforms is that the tax-interaction effect tends to be larger in absolute magnitude than the revenue-recycling effect; hence, revenue-neutral reforms are more costly in a second-best setting than in a first-best world. The intuition behind this result is that environmental taxes generally are more narrow than factor taxes; hence, they are less efficient mechanisms for raising revenue (or more costly in terms of consumption of non-environmental goods) than factor taxes are. The very characteristics of environmental taxes that make them attractive for achieving environmental goals – namely, their focus on particular, pollution-generating activities or processes – make them unattractive as instruments for raising revenue.

The dominance of the tax-interaction effect over the revenue-recycling effect bears on the double-dividend claim about revenue-neutral reforms. The double (i.e., second) dividend arises only if the costs of revenue-neutral environmental reforms are zero or negative, that is, if such reforms reduce the overall gross costs of the tax system. This is an even stronger requirement than the requirement that revenue-neutral reforms be less costly in a second-best world than in a first-best setting. Thus, *a fortiori,* the dominance of the tax-interaction effect over the revenue-recycling effect refutes the double-dividend argument. An important caveat is in order, however. As we have discussed, if the existing tax system is highly inefficient along other, non-environmental dimensions (for example, if capital is excessively taxed relative to labor), there may be scope for the double dividend after all. A double dividend is possible if "green" tax reform helps eliminate pre-existing inefficiencies of this sort. The question arises as to why green tax reform is necessary to deal with these inefficiencies, since in

principle they could be addressed more directly through "ordinary" tax reform. This raises difficult political issues that lie beyond the scope of this chapter.

6.3. Instrument choice

Tax interactions are also crucial to the choice between environmental taxes and other, non-tax instruments for environmental protection. Non-auctioned pollution quotas produce the same costly tax-interaction effect that environmental taxes do. But, in contrast with environmental taxes whose revenues are used to finance cuts in distortionary taxes, such quotas fail to enjoy the beneficial (in efficiency terms) revenue-recycling effect. As we have seen, the absence of the revenue-recycling effect puts quotas at a significant efficiency disadvantage: the net efficiency gains (incorporating environmental benefits) from quotas may be much lower than those under environmental taxes. Indeed, in some circumstances the inability to exploit the revenue-recycling effect may make it impossible to generate efficiency improvements through non-auctioned pollution quotas.

The presence of uncertainty about abatement costs and benefits, and the associated costs of monitoring and enforcement, complicate the problem of instrument choice. Once we account for these issues, the efficiency ranking of taxes, quotas, and other instruments (such as performance standards and mandated technologies) becomes less clear. As we have noted, much depends on the nature of the uncertainty and the monitoring and enforcement costs. These complications can at least partly explain why policy makers often have persisted in favoring command-and-control approaches over incentive-based policies.

6.4. Distributional issues

Distributional considerations also complicate instrument choice. Quotas and taxes differ in their distributional impacts, and one of the potential attractions of non-auctioned quotas is that they involve a smaller transfer of wealth from polluters[75] to taxpayers. This distributional aspect has powerful political implications, and helps explain why the political process tends to favor grandfathered permits over auctioned permits or taxes. These distributional attractions need to be weighed against the efficiency disadvantages of quotas. The present chapter could not provide the weights to be assigned to these competing goals, but it was able to clarify the efficiency cost of meeting the distributional objectives.

Some authors have attempted to examine jointly the distributional and efficiency issues. They consider, in particular, how the government might employ a nonlinear tax system to meet all of its distributional objectives (including ironing out the

[75] Or, more precisely, the owners, workers, and consumers that ultimately bear the burden of pollution taxes.

distributional effects of environmental public goods), and employ environmental taxes to serve the goal of providing the optimal amount of environmental quality. If the government has access to a nonlinear tax and if other, special conditions obtain, the Pigouvian rule for optimal environmental taxation can apply after all. These conditions are unlikely to prevail in the real world, however. Generally, it is unrealistic to expect that distributional consequences will be ironed out through adjustments to a nonlinear income tax. This means that the Pigouvian rule usually will not apply and that distributional consequences of environmental policies have to be accounted for in the choice of environmental policy instruments.

6.5. Areas for future research

Although no one can predict with certainty the returns from academic research, the discussion above suggests to us some areas where further research explorations might yield significant payoffs. The analysis of environmental taxation often lacks attention to real-world complications, so that the prescribed remedies become irrelevant to policy discussions. Two key complications are information problems and associated implementation issues; a closer attention to these complications might bring substantial rewards. In the past two decades, progress on the design of marketable pollution-permits programs and on deposit-refund systems proved to be very useful to policymakers: tradeable permits helped overcome significant information burdens encountered by regulators in the face of heterogenous producers, and deposit–refund systems helped overcome significant monitoring problems that sometimes bedeviled environmental taxes. Unfortunately, information, monitoring, and enforcement problems persist in many areas where tradeable-permits or deposit–refund systems are not feasible. New instruments are needed to deal with these problems.

Another key difficulty with current work is that, too often, efficiency assessments are made in isolation, without attention to distributional impacts. Distributional considerations carry a great deal of political force, and thus studies that integrate efficiency and distributional assessments seem especially valuable. A major challenge to environmental policy making seems to be the design of policies that achieve efficiency goals without producing unacceptable distributional outcomes. Public-economics textbooks often suggest that distributional activities should be carried out only by the "distribution arm" of the public sector, leaving regulatory authorities free to concentrate exclusively on efficiency. This separation of functions is intellectually appealing, but unfortunately the political process does not seem to allow such separation of impacts when policy proposals are debated. This suggests a value to research that helps design transfer mechanisms to accompany environmental policies that otherwise would have undesirable distributional consequences.

References

Adar, A., and J.M. Griffin (1976), "Uncertainty and the choice of pollution control instruments", Journal of Environmental and Economics Management 3:178–188.

Aronsson, T. (1999), "On cost benefit rules for green taxes", Environmental and Resource Economics 13:31–43.

Atkinson, A.B., and N.H. Stern (1974), "Pigou, taxation, and public goods", Review of Economic Studies 41:119–128.

Atkinson, A.B., and J.E. Stiglitz (1976), "The design of tax structure: direct versus indirect taxation", Journal of Public Economics 6:55–75.

Auerbach, A.J., and J.R. Hines Jr (2001), "Taxation and economic efficiency", in: Alan J. Auerbach and Martin Feldstein, eds., Handbook of Public Economics, Vol. 3 (Elsevier, Amsterdam) ch. 21, this volume.

Ballard, C.L., and D. Fullerton (1992), "Distortionary taxes and the provision of public goods", Journal of Economic Perspectives 6(3):117–131.

Ballard, C.L., and S.G. Medema (1993), "The marginal efficiency effects of taxes and subsidies in the presence of externalities", Journal of Public Economics 52:199–216.

Ballard, C.L., J.B. Shoven and J. Whalley (1985), "General equilibrium computations of the marginal welfare costs of taxes in the United States", American Economic Review 77:11–23.

Baumol, W.J., and W.E. Oates (1988), The Theory of Environmental Policy, 2nd Edition (Cambridge University Press, Cambridge, UK).

Boadway, R., and M. Keen (1993), "Public goods, self-selection and optimal income taxation", International Economic Review 34:463–478.

Boadway, R., and M. Marchand (1995), "The use of public expenditures for redistributive purposes", Oxford Economic Papers 47:45–59.

Bohm, P. (1981), Deposit–Refund Systems: Theory and Applications to Environmental, Conservation, and Consumer Policy (The Johns Hopkins University Press, Baltimore, London).

Bovenberg, A.L., and R.A. de Mooij (1994a), "Environmental levies and distortionary taxation", American Economic Review 84(4):1085–1089.

Bovenberg, A.L., and R.A. de Mooij (1994b), "Environmental taxes and labour-market distortions", European Journal of Political Economy 10(4):655–684.

Bovenberg, A.L., and L.H. Goulder (1996), "Optimal environmental taxation in the presence of other taxes: general equilibrium analyses", American Economic Review 86(4):985–1000.

Bovenberg, A.L., and L.H. Goulder (1997), "Costs of environmentally motivated taxes in the presence of other taxes: general equilibrium analyses", National Tax Journal 70(1):59–87.

Bovenberg, A.L., and L.H. Goulder (2001), "Neutralizing the adverse industry impacts of CO_2 abatement policies: what does it cost? in: Carlo Carraro and Gilbert Metcalf, eds., Behavioral and Distributional Impacts of Environmental Policies (University of Chicago Press).

Bovenberg, A.L., and F. van der Ploeg (1994a), "Green policies in a small open economy", Scandinavian Journal of Economics 96(3):343–363.

Bovenberg, A.L., and F. van der Ploeg (1994b), "Environmental policy, public finance and the labour market in a second-best world", Journal of Public Economics 55:349–370.

Bovenberg, A.L., and F. van der Ploeg (1998a), "Consequences of environmental tax reform for involuntary unemployment and welfare", Environmental and Resource Economics 12:137–150.

Bovenberg, A.L., and F. van der Ploeg (1998b), "Tax reform, structural unemployment and the environment", Scandinavian Journal of Economics 100(3):593–610.

Browning, E.K. (1987), "On the marginal welfare cost of taxation." American Economic Review 77(1):11–23.

Brunello, G. (1996), "Labour market institutions and the double dividend hypothesis: an application of the WARM model", in: C. Carraro and D. Siniscalco, eds., Environmental Fiscal Reform and Unemployment (Kluwer, Dordrecht) pp. 139–170.

Buchanan, J.M., and G. Tullock (1975), "Polluters profits and political response: direct controls versus taxes", American Economic Review 65:139–147.

Capros, P., P. Georgakopoulos, S. Zografakis, S. Proost, D. van Regemorter, K. Conrad, T. Schmidt, Y. Smeers and E. Michiels (1996), "Double dividend analysis: first results of a general equilibrium

model (GEM-E3) linking the EU-12 countries", in: C. Carraro and D. Siniscalco, eds., Environmental Fiscal Reform and Unemployment (Kluwer, Dordrecht) pp. 193–228.
Carraro, C., and D. Siniscalco, eds (1996), Environmental Fiscal Reform and Unemployment (Kluwer, Dordrecht).
Christiansen, V. (1981), "Evaluation of public projects under optimal taxation", Review of Economic Studies 48:447–457.
Christiansen, V. (1996), "Green taxes: a note on the double dividend and the optimum tax rate", CES Working Paper 107 (University of Munich).
Coase, R.H. (1960), "The problem of social cost", Journal of Law and Economics 3:1–44.
Corlett, W.J., and D.C. Hague (1953), "Complementarity and the excess burden of taxation", Review of Economic Studies 21:21–30.
Cornes, R. (1980), "External effects: an alternative formulation", European Economic Review 14: 307–321.
Cremer, H., F. Gahvari and N. Ladoux (1998), "Externalities and optimal taxation", Journal of Public Economics 70(3):343–364.
Crocker, T.D. (1966), "The structuring of atmospheric pollution control systems." in: Harold Wolozin, ed., The Economics of Air Pollution (Norton, New York) pp. 61–86.
Dales, J. (1968), Pollution, Property and Prices (University Press, Toronto).
Deaton, A. (1977), "Equity, efficiency and the structure of indirect taxation", Journal of Public Economics 8:299–312.
Diamond, P.A., and J.A. Mirrlees (1971), "Optimal taxation and public production I: production efficiency; and II: tax rules", American Economic Review 61.
Downing, P.B., and W.D. Watson Jr (1974), "The economics of enforcing air pollution controls", Journal of Environmental Economics and Management 1(3):219–236.
Eskeland, G.S. (2000), "Externalities and production efficiency", Working Paper (The World Bank).
Eskeland, G.S., and S. Devarajan (1995), "Taxing bads by taxing goods: towards efficient pollution control with presumptive charges." in: Lans Bovenberg and Sijbren Cnossen, eds., Public Economics and the Environment in an Imperfect World (Kluwer, Dordrecht) pp. 61–112.
Farzin, H., and O. Tahvonen (1996), "Global carbon cycle and the optimal time path of a carbon tax", Oxford Economic Papers 48(4):515–536.
Feldstein, M. (1976), "On the theory of tax reform", Journal of Public Economics 6:77–104.
Fishelson, G. (1976), "Emission control policies under uncertainty", Journal of Environmental Economics and Management 3:189–197.
Fullerton, D. (1996), "Why have separate environmental taxes?" in: James M. Poterba, ed., Tax Policy and the Economy 10 (MIT Press, Cambridge, MA.).
Fullerton, D. (1997), "Environmental levies and distortionary taxation: comment", American Economic Review 87.
Fullerton, D., and G.E. Metcalf (2001), "Environmental controls, scarcity rents, and pre-existing distortions", Journal of Public Economics 80(2):249–267.
Fullerton, D., and A. Wolverton (1997), "The case for a two-part instrument: presumptive tax and environmental subsidy", Working Paper 5993 (NBER, Cambridge, MA).
Gaube, T. (1998), "Distortionary taxes preserve the environment." Discussion Paper A-579 (Department of Economics, University of Bonn).
Goulder, L.H. (1994), "Energy taxes: traditional efficiency effects and environmental implications", in: James M. Poterba, ed., Tax Policy and the Economy 8 (MIT Press, Cambridge, MA) pp. 105–158.
Goulder, L.H. (1995a), "Environmental taxation and the 'double dividend:' a reader's guide", International Tax and Public Finance 2(2):157–183.
Goulder, L.H. (1995b), "Effects of carbon taxes in an economy with prior tax distortions: an intertemporal general equilibrium analysis", Journal of Environmental Economics and Management 29:271–297.
Goulder, L.H., and K. Mathai (2000), "Optimal CO_2 abatement in the presence of induced technological change", Journal of Environmental Economics and Management 39(1):1–38.

Goulder, L.H., I.W.H. Parry and D. Burtraw (1997), "Revenue-raising vs. other approaches to environmental protection: the critical significance of pre-existing tax distortions", RAND Journal of Economics 28(4):708–731.

Goulder, L.H., I.W.H. Parry, R.C. Williams III and D. Burtraw (1999), "The cost-effectiveness of alternative instruments for environmental protection in a second-best setting", Journal of Public Economics 72(3):329–360.

Hahn, R.W. (1984), "Market power and transferable property rights", Quarterly Journal of Economics 99:753–765.

Hahn, R.W. (1989), "Economic prescriptions for environmental problems: how the patient followed the doctor's orders." Journal of Economic Perspectives 3:103–110.

Hahn, R.W., and R. Noll (1982), "Designing a market for tradeable permits", in: Wesley Magat, ed., Reform of Environmental Regulation (Ballinger, Cambridge) pp. 119–146.

Hansson, I., and C. Stuart (1985), "Tax revenue and the marginal cost of public funds in Sweden", Journal of Public Economics 27(3):331–353.

Harford, J.D. (1978), "Firm behavior under imperfectly enforceable pollution standards and taxes", Journal of Environmental Economics and Management 5(1):26–43.

Harrington, W. (1988), "Enforcement leverage when penalties are restricted", Journal of Public Economics 37:29–53.

Jorgenson, D.W., and P.J. Wilcoxen (1990), "Environmental regulation and U.S. economic growth", The Rand Journal of Economics 21(2):314–340.

Jorgenson, D.W., and P.J. Wilcoxen (1996), "Reducing U.S. carbon emissions: an econometric general equilibrium assessment", in: Darius Gaskins and John Weyant, eds., Reducing Global Carbon Dioxide Emissions: Costs and Policy Options Energy Modeling Forum (Stanford University, Stanford, CA).

Kaplow, L. (1996), "A note on the optimal supply of public goods and the distortionary cost of taxation", National Tax Journal 51:117–125.

Kaplow, L., and S. Shavell (1997), "On the superiority of corrective taxes to quantity regulation", Working paper (Harvard Law School, Harvard University, Cambridge, MA).

Kaplow, L., and S. Shavell (2001), "Economic analysis of law", in: Alan J. Auerbach and Martin Feldstein, eds., Handbook of Public Economics, Vol. 3 (Elsevier, Amsterdam) ch. 25.

Koskela, E., and R. Schöb (1999), "Alleviating unemployment: the case for green tax reforms", European Economic Review 43:1723–1746.

Layard, R., S. Nickell and R. Jackman (1991), Unemployment. Macroeconomic Performance and the Labor Market (Oxford University Press, Oxford).

Lee, D.R., and W.S. Misiolek (1986), "Substituting pollution taxation for general taxation: some implications for efficiency in pollution taxation", Journal of Environmental Economics and Management 13:338–347.

Lerner, A.P. (1971), "The 1971 report of the Presidents' Council of Economic Advisers: priorities and efficiency", American Economic Review 61:527–530.

Lewis, T. (1996), "Protecting the environment when costs and benefits are privately known", RAND Journal of Economics 27(4):819–847.

Manne, A.S., and R.G. Richels (1992), Buying Greenhouse Insurance: The Economic Costs of CO_2 Emissions Limits (MIT Press, Cambridge, MA).

Menell, P. (1990), "Beyond the throwaway society: an incentive approach to regulating municipal solid waste", Ecology Law Quarterly 17:655.

Metcalf, G.E. (2000), "Environmental levies and distortionary taxation: Pigou, taxation, and pollution", Working Paper 7917 (National Bureau of Economic Research).

Mirrlees, J.A. (1976), "Optimal tax theory: a synthesis", Journal of Public Economics 6:327–358.

Misiolek, W.S., and H.W. Elder (1989), "Exclusionary manipulation of markets for pollution rights", Journal of Environmental Economics and Management 16:156–166.

Montgomery, W.D. (1972), "Markets in licenses and efficient pollution control programs", Journal of Economic Theory 5:395–418.

Newell, R., and W. Pizer (1998), "Regulating stock externalities under uncertainty", Discussion Paper 99-10 (Resources for the Future).

Ng, Y.K. (1980), "Optimal corrective taxes or subsidies when revenue-raising imposes an excess burden", American Economic Review 70:744–751.

Nielsen, S.B., L.H. Pedersen and P.B. Sørensen (1995), "Environmental policy, pollution, unemployment and endogenous growth", International Tax and Public Finance 2:185–205.

Nordhaus, W.D. (1980), "Thinking about carbon dioxide: theoretical and empirical aspects of optimal growth strategies", Discussion Paper 565 (Cowles Foundation, Yale University).

Nordhaus, W.D. (1982), "How fast should we graze the global commons?" American Economic Review (Papers and Proceedings) 72(2):242–246.

Nordhaus, W.D. (1994), Managing the Global Commons: The Economics of Climate Change (MIT Press, Cambridge, MA).

Oates, W.E. (1993), "Pollution charges as a source of public revenues", in: Herbert Giersch, ed., Economic Progress and Environmental Concerns 135-52 (Springer, Berlin).

Oates, W.E. (1995), "Green taxes: can we protect the environment and improve the tax system at the same time?" Southern Economic Journal 61(4):914–922.

Parry, I.W.H. (1995), "Pollution taxes and revenue recycling", Journal of Environmental Economics and Management 29:S64–S77.

Parry, I.W.H. (1997), "Environmental taxes and quotas in the presence of distorting taxes in factor markets", Resource and Energy Economics 19:203–220.

Parry, I.W.H. (1998), "A second-best analysis of environmental subsidies", International Tax and Public Finance 5:153–170.

Parry, I.W.H., and A. Bento (2000), "Tax deductions, environmental policy, and the 'double dividend' hypothesis", Journal of Environmental Economics and Management 39(1):67–96.

Parry, I.W.H., R.C. Williams III and L.H. Goulder (1999), "When can carbon abatement policies increase welfare? The fundamental role of distorted factor markets", Journal of Environmental Economics and Management 37:52–84.

Peck, S.C., and T.J. Teisberg (1994), "Optimal carbon emissions trajectories when damages depend on the rate or level of global warming", Climatic Change 28:289–314.

Peck, S.C., and Y.S. Wan (1996), "Analytical solutions of simple optimal greenhouse gas emission models", Working paper (Electric Power Research Institute, Palo Alto, CA).

Pezzey, J. (1992), "The symmetry between controlling pollution by price and controlling it by quantity", Canadian Journal of Economics 25:983–999.

Pigou, A.C. (1938), The Economics of Welfare, Fourth Edition (Weidenfeld and Nicolson, London).

Pirttilä, J., and M. Tuomala (1997), "Income tax, commodity tax, and environmental policy", International Tax and Public Finance 4:379–393.

Poterba, J.M. (1993), "Global warming: a public finance perspective", Journal of Economic Perspectives 7(4):47–63.

Proost, S., and D. van Regemorter (1995), "The double dividend and the role of inequality aversion and macroeconomic regimes", International Tax and Public Finance 2(2):207–220.

Roberts, M.J., and M. Spence (1976), "Effluent charges and licences under uncertainty", Journal of Public Economics 5:193–208.

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

Sandmo, A. (1975), "Optimal taxation in the presence of externalities", Swedish Journal of Economics 77:86–98.

Sandmo, A. (1980), "Anomaly and stability in the theory of externalities", Quarterly Journal of Economics 94:799–807.

Schmutzler, A., and L.H. Goulder (1997), "The choice between emissions taxes and output taxes under imperfect monitoring", Journal of Environmental Economics and Management 32:51–64.

Schneider, K. (1997), "Involuntary unemployment and environmental policy: the double dividend hypothesis", Scandinavian Journal of Economics 99:45–59.
Schöb, R. (1996), "Choosing the right instrument", Environmental and Resource Economics 8:399–416.
Shackleton, R., M. Shelby, A. Cristofaro, R. Brinner, J. Yanchar, L.H. Goulder, D.W. Jorgenson, P.J. Wilcoxen and P. Pauly (1996), "The efficiency value of carbon tax revenues", in: Darius Gaskins and John Weyant, eds., Reducing Global Carbon Dioxide Emissions: Costs and Policy Options (Energy Modeling Forum, Stanford, CA).
Shah, A., and B. Larsen (1992), "Carbon taxes, the greenhouse effect and developing countries", World Bank Policy Research Working Paper Series No. 957 (The World Bank, Washington, D.C.).
Sinclair, P.J.N. (1994), "On the optimum trend of fossil fuel taxation", Oxford Economic Papers 46:869–877.
Slemrod, J., and S. Yitzhaki (2000), "Integrating expenditure and tax decisions: the marginal cost of funds and the marginal benefit of projects", Working paper (Department of Economics, Hebrew University).
Stavins, R.N. (1995), "Transaction costs and tradeable permits", Journal of Environmental Economics and Management 29:133–147.
Stavins, R.N. (1996), "Correlated uncertainty and policy instrument choice", Journal of Environmental Economics and Management 30:233–253.
Stavins, R.N. (2000), "Market-based environmental policies", in: Paul R. Portney and Robert N. Stavins, eds., Public Policies for Environmental Protection (Resources for the Future, Washington, D.C.) pp. 31–76.
Swierzbinski, J. (1994), "Guilty until proven innocent – regulation with costly and limited enforcement", Journal of Environmental Economics and Management 25:127–146.
Terkla, D. (1984), "The efficiency value of effluent tax revenues", Journal of Environmental Economics and Management 11:107–123.
Tietenberg, T. (1985), Emissions Trading: An Exercise in Reforming Pollution Policy (Resources for the Future, Washington, DC).
Tietenberg, T. (1997), "Tradeable permits and the control of air pollution in the United States", Paper prepared for the 10^{th} Anniversary Jubilee Edition of Zeitschrift Für Angewandte Umweltforschung.
Ulph, A., and D. Ulph (1994), "The optimal time path of a carbon tax", Oxford Economic Papers 46:857–868.
Upton, C.W. (1971), "The allocation of pollution rights", Urban Economics Report 59 (University of Chicago).
US Environmental Protection Agency (1996), National Air Quality and Emissions Trends Report, EPA Document Number 454/R-97-013 (US Environmental Protection Agency).
Weitzman, M.L. (1974), "Prices vs. quantities", Review of Economic Studies 41:477–491.
Wijkander, H. (1985), "Correcting externalities through taxes on-subsidies to related goods", Journal of Public Economics 28(1):111–25.
Wildasin, D.E. (1984), "On public good provision with distortionary taxation", Economic Inquiry 22:227–243.
Williams III, R.C. (1997), "Environmental tax interactions when environmental quality has productivity or health effects", Working paper (Stanford University).

Chapter 24

POLITICAL ECONOMICS AND PUBLIC FINANCE*

TORSTEN PERSSON

Institute for International Economic Studies, Stockholm University, Stockholm, Sweden

GUIDO TABELLINI

IGIER, Bocconi University, Milano, Italy

Contents

* We thank a number of colleagues – especially Alan Auerbach, David Baron, Tim Besley, Francesco Daveri, Avinash Dixit, and Gerard Roland – and Ph.D. students – especially Gisela Waisman – for comments on earlier drafts. Christina Lönnblad and Alessandra Startari provided editorial assistance. The Bank of Sweden Tercentenary Foundation, the European Commission (a TMR-Grant), and Bocconi University supported the underlying research.

Handbook of Public Economics, Volume 3, Edited by A.J. Auerbach and M. Feldstein

Abstract

Observed fiscal policy varies greatly across time and countries. How can we explain this variation? This paper surveys the recent literature that has tried to answer this question. We adopt a unified approach in portraying public policy as the equilibrium outcome of an explicitly specified political process. We divide the material into three parts. In Part I, we focus on median-voter equilibria that apply to policy issues where disagreement between voters is likely to be one-dimensional. We thus study the general redistributive programs, typical of the modern welfare state: redistribution between rich and poor, young and old, employed and unemployed, and labor and capital. In Part II we study special-interest politics. Here, the policy problem is multi-dimensional and we focus on specific political mechanisms: we study legislative bargaining, lobbying, and electoral competition, as well as the possible interactions between these different forms of political activity. Finally, in Part III we deal with "comparative politics", namely policy choice under alternative political constitutions. Here, we model the rationale for separation of powers; we also contrast stylized features of majoritarian and proportional electoral rules, as well as congressional and parliamentary political regimes, focusing on their implications for rent extraction by politicians, redistribution and public goods provision.

Keywords

elections, interest groups, agency problems, welfare-state programs, comparative politics

JEL classification: D7, E6, H0

1. General introduction

Observed fiscal policy varies greatly across time and countries. Over time, the size of government has grown in a striking way. In 14 OECD countries, for which data are available, average government spending was less than 10% of GDP just before World War I. It had doubled to 18% just before World War II. By 1960, it was close to 30%. And by the mid 1990s, it had reached almost 50%. The growth of government accelerated after the mid 1930s, and slowed down towards the late 1980s. Equally striking are the differences across countries. In the late 1990s, total government spending as a fraction of GDP stood at about 60% in Sweden, and well above 50% in many countries of continental Europe, but around 35% in Japan, Switzerland and the USA.

The composition of spending also varies greatly across time and countries. Government transfers is the component that accelerated most rapidly: in 1937 transfers amounted to only 4% of GDP, on average, in 7 OECD countries for which data are available; by the 1990s, they had reached over 20%. Over the same period, government consumption also increased, but by less (8% of GDP). Public investment, in contrast, has remained roughly constant since 1970, at around 3% of GDP, in most countries. Moreover, the big spenders with regard to public consumption are not always the countries with large governments: in the 1990s, the USA and the UK had higher government consumption than the average of 17 OECD countries, even though their total government spending was considerably smaller than the average of the same countries. The cross-country variation in size and composition of spending is even greater in a larger set of more heterogeneous countries, also including developing countries. Finally, the quality and effectiveness of government activities vary considerably across countries, even among countries at comparable levels of development[1].

How can we explain such variation across time and countries? Is it associated with systematic variation in other aspects of economic policy? What is the role of alternative political constitutions and collective choice procedures in explaining fiscal policy outcomes? Are the observed patterns of spending and taxation likely to reflect socially optimal policy choices – given some normative criterion? If not, how can we account for the deviations from the normative benchmark? Do these deviations reflect the wishes of a majority of the voters? These fundamental questions were raised long ago by researchers in the so-called public-choice school. But it is fair to say that until recently, they have been neglected by traditional economic analysis. Specifically, policy analysis in traditional public finance was almost entirely normative, ignoring the positive theory of policy choice. This is no longer so. A growing body of research

[1] Tanzi and Schuknecht (1995) discuss historical data on government spending for OECD countries, while Persson and Tabellini (1999b, 2001) consider larger groups of countries. The quality of government activities in a very broad group of countries is discussed empirically in La Porta et al. (1999).

now tackles positive public-finance questions head on, fruitfully combining economic and political analysis. The goal of this chapter is to provide a selective survey of this emerging literature.

We try to look ahead, at the most promising areas of new research in this emerging literature on *political economics*. In the process, we do not always give full justice to the earlier literature on similar issues. One reason is the excellent survey by Inman (1987), in an earlier volume of this Handbook, which gives a general account of the literature up until the early 1980s. In particular, Inman shows how the literature on political economy relates to some of the fundamental results in social choice and philosophy. There are also excellent surveys of the public-choice approach to economic policy; see, in particular, Frey (1983) and Mueller (1989, 1997). Another reason is that some of the earlier literature was based on spatial models of voting, where individual preferences for public policy were not based on explicit economic models. In this survey, instead, we always combine economic theory with the analysis of alternative collective choice procedures. Even though we sometimes study very simple model economies, the main goal – for each topic that we cover – is always to explain some specific economic policy outcomes.

In terms of recent textbook treatments of the field, this survey substantially overlaps with part of our own [Persson and Tabellini (2000)]. Drazen (2000) also covers some of the topics we deal with here, although his main focus is on macroeconomic issues. Grossman and Helpman (2001) deal in depth with many of the issues concerning special-interest politics, covered in Part II of this survey. To avoid overlap with other existing surveys, we do not discuss the literatures on local public finance, macroeconomic policy, trade and international economic policy. We also restrict ourselves to *static* models of public finance, or more precisely, models with one-time policy choices[2].

We adopt a unified approach in portraying public policy as the equilibrium outcome of an explicitly specified political process. Policy choices are not made by a hypothetical benevolent social planner, but by purposeful and rational political agents participating in a well-defined decision-making process. Alternative theories seek to capture different features of political institutions and alternative modes of political behavior. Even though there is a variety of models, some general determinants of economic policies emerge from the analysis.

Public policy must strike a balance between the *conflicting interests of different voters*. The conflict largely reflects socio-economic factors, deriving from differences in income, age, employment status, geographical residence, occupation, or the like. In the simplest setting, these socio-economic factors shape the distribution of voters' policy preferences, which, in turn, get aggregated into public policy by the majority principle.

[2] Scotchmer (2002), Persson and Tabellini (1995, 1999a), Dixit (1996a), Inman and Rubinfeld (1997) and Rodrik (1995) survey these other topics.

But the resolution of conflicting interests also reflects *political power.* In some cases, the determinants of political power are obvious. For instance, redistribution harms individuals unrepresented or under-represented in the political process, like future generations or citizens not organized in a political lobby. In other cases, political power derives from less obvious features of the political process. For instance, ideologically neutral and well-informed voters are more influential, because they often become the arbiter of the electoral competition between vote-maximizing parties. Political power is particularly important when it comes to special-interest politics: concentration of benefits and dispersion of costs create very uneven incentives for trying to influence public policy. The groups benefitting most from the policy have strong incentives to get organized and build political power, at the expense of everyone else. This distorts the policymaker's incentives and leads to biased equilibrium outcomes, including distorted allocations or large government spending. This idea is familiar from the early public-choice literature. More recent contributions have studied "structural" models where policy outcomes are suboptimal, even if political decision making is centralized, as long as groups or individuals acting in a decentralized fashion retain political influence.

In representative democracies, public policy must also strike a balance between the *conflicting interests of voters and politicians.* This prospective *agency* problem is also an old theme of the public-choice school. Sometimes, the problem is challenged by the argument that electoral competition between vote-maximizing candidates could remove the source of inefficiency: if there is an inefficient status quo, what prevents a vote-maximizing political entrepreneur from running as a candidate, promising efficient policies?[3] When politicians cannot commit to enforceable or verifiable state-contingent electoral promises, however, the benefits of political competition are weakened and some agency rents remain. The struggle to capture those rents affects the policy outcome. For instance, elected officials may have an incentive to expand tax revenues, since that makes it easier to reap rents from office.

But we also encounter *conflicting interests of different politicians* about how to split available rents. The resolution of these conflicts hinges on the constitution, as the details of the decision-making procedure determine who has the power to exploit the political rents for his own benefit. Different constitutions may also give the voters more or less control over their elected politicians by holding them accountable in general elections.

We divide the material into three parts. The division partly reflects methodology, partly substance. In Part I, we focus on median-voter equilibria that apply to policy issues where disagreement between voters is likely to be one-dimensional. As the political mechanism is so simple, we can add more economic structure. We thus study the *general redistributive programs,* typical of the modern welfare state. Specifically, we deal with redistribution between rich and poor, young and old,

[3] Stigler (1971) and Wittman (1989) e.g. argue along these lines.

employed and unemployed, and labor and capital. We can think of these median-voter equilibria as implemented in Downsian electoral competition between vote-maximizing candidates (parties). But all the equilibria in Part I are *preference induced,* in that they only depend on the distribution of individual policy preferences.

For many aspects of public finance, however, the simplification to policy conflict along a single dimension is too hard to swallow. More general, multi-dimensional policies generate more narrowly defined special interests. Such situations require precise institutional assumptions to overcome the problems posed by Arrow's impossibility theorem. We illustrate a number of possibilities in Part II, on *special-interest politics.* Here, our approach is opposite to that in Part I. Thus, we simplify on the economic front, by studying a common problem of local, group-specific, public goods provision. This way, we illustrate a number of alternative approaches for analyzing how the resulting policy conflict may be resolved. Each of these approaches highlights a different aspect of the political process and therefore suggests different determinants of which groups will gain and which will lose. Specifically, we study legislative bargaining, lobbying, and electoral competition, as well as the possible interactions between these different forms of political activity. All the equilibria in Part II are *structure induced,* in that they crucially depend on the assumed institutions.

Finally, Part III deals with a set of questions brought together under the label of *comparative politics,* as we deal with policy choice under alternative political constitutions. Here, we explicitly view a political constitution as an incomplete contract. Politicians cannot commit to verifiable state-contingent electoral promises, which aggravates the agency problems between voters and elected representatives. According to this approach, the reason why different constitutions may produce systematically different policy choices is that they entail different allocations of control rights to politicians and voters. We illustrate some key ideas in this nascent literature, drawing on the results in previous parts. Specifically, we model some stylized features of different electoral rules, such as district magnitude and the electoral formula. We discuss how these rules shape the trade-off between broad public-policy programs and redistributive programs targeted to narrow groups, as well as their effectiveness in containing rent extraction: We also compare key features of different political regimes. Specifically, we juxtapose congressional and parliamentary regimes, arguing that their different allocations of proposal and veto rights have important consequences for how well voters can control rent extraction by their political representatives, and for how redistribution and public-goods provision are traded off in the legislative process.

Each of the three parts starts with a general introduction providing a more detailed road map to the following sections. We typically give references to the key contributions on which we build at the beginning of each section and in connection with the main results. More extensive references are instead collected at the end of each section in special subsections labeled "Notes on the literature".

Part I. General redistributive politics

One of the prime goals of political economics is to study the policy implications of conflicting interests among individual citizens. In this first part we study conflict and heterogeneity in the simplest possible political set-up, where political equilibria exclusively reflect the *preferences* of the citizens. Except in the very last section, we always study simple models of electoral competition, where two office-motivated candidates who only care about winning make binding policy announcements ahead of the elections. The goal of the analysis is to understand why different economic agents have different policy preferences, how these preferences can be derived from the economic role of individuals, and how they shape economic policy in political equilibrium. Given the simplicity of the political mechanism, the value added of the analysis lies in the derivation of individual policy preferences from a well-specified economic environment.

Two models of electoral competition have been proposed in the literature. The simplest and most well-known is the Downsian median-voter model [Downs (1957)], where the candidates are identical and voters only care about the announced policies. An equilibrium policy is a so-called *Condorcet winner;* that is, a policy that cannot be beaten by any other policy in a pair-wise majority vote. Such policies only exist under restrictive conditions on voters' preferences; the classical condition requires voters' preferences to be single-peaked. Even though we can allow slightly more general conditions, we must essentially assume that the political problem is one-dimensional – either because the policy space itself is one-dimensional, or because voters' preferences over a multi-dimensional policy are smooth enough to allow their disagreement to be projected on a single-dimensional space[4]. If these conditions are satisfied, the equilibrium of the electoral competition game has candidates converging to the policy preferred by the voter with median preferences. For this reason, such equilibria are often called *median-voter equilibria.* From a positive point of view, the equilibrium policy reflects the *distribution* of policy preferences in the population, but not the intensity of such preferences. Voters with very strong policy preferences influence the equilibrium just as much as voters who are almost indifferent.

Median-voter equilibria have been extensively studied by economists. One explanation is undoubtedly ease of analysis: such equilibria constitute the solution to an optimal taxation problem, given a very special social welfare function, where only the utility of the median individual carries positive weight. The simple political setting has the virtue of enabling the researcher to study rich policy problems associated with quite complex economic environments. Another explanation for their popularity may be that median-voter equilibria identify some of the basic political forces shaping economic

[4] The literature includes a number of generalizations of the single-peakedness condition. The main ones are the monotonicity condition of Roberts (1977), the intermediate-preference condition of Grandmont (1978), and the order-restricted preferences of Rothstein (1990). Gans and Smart (1996) formulate a useful single-crossing condition, which incorporates many other restrictions as special cases.

policy. Virtually everyone dislikes the equilibrium policy. But half the electorate wants to move policy in one direction, and the other half wants to move it in the opposite direction. In this sense, a median-voter optimum resembles a Walrasian equilibrium: once we have reached an equilibrium, fundamental forces tend to keep policy in place.

The second model of electoral competition studied in the literature is the *probabilistic voting* model [Enelow and Hinich (1982), Ledyard (1984), Lindbeck and Weibull (1987)]. Here voter behavior includes a random component and voters with more intense policy preferences are more likely to vote in favor of the preferred policy. This assumption can be interpreted as reflecting voting costs, which make almost indifferent voters more likely to abstain. An alternative interpretation is that voters have policy preferences as well as intrinsic preferences over candidates, and the latter are more likely to be dominated when the former are stronger. The equilibrium policy can be computed as the solution to a modified social welfare function, where individual weights reflect the probability that citizens reward policy favors with their vote. Thus, groups with higher turnout rates or with weaker preferences over candidates (as opposed to policies) are more influential in shaping equilibrium policy.

This model of electoral competition has been used less often in political economics, perhaps because it is less well known. Compared to the median-voter model it has advantages and drawbacks. One obvious advantage is that a probabilistic voting equilibrium exists under much more general circumstances than a Condorcet winner even when policy conflict is multidimensional. A drawback is that economic theory has very little to say about a central determinant of the equilibrium, namely the probability of voting in favor of the preferred policy. The median-voter model is more self-contained: the economic environment shapes policy preferences and these in turn uniquely pin down the political equilibrium. In a probabilistic voting model, instead, we need to look beyond economics, and ask which groups are more likely to abstain, or to have strong preferences over political candidates. While the model identifies a potentially important and plausible political force, namely the voting behavior of different groups, it has less to say about what drives this behavior.

We do not present the probabilistic voting model until Part II, however, and instead rely on the median-voter model for most of this Part. One reason for this choice is that we study broad redistributive programs, such as those typical of the welfare state, that are *not* narrowly designed to target small groups of beneficiaries. Conflict over these programs is typically aligned over a few dimensions, since the programs redistribute between broad socio-economic groups. Hence, the assumptions needed to guarantee existence of the median-voter equilibrium are more likely to be satisfied, compared to the multidimensional policy problems studied in later parts of the chapter.

Specifically, Section 2 studies redistribution between *rich and poor* voters. Here, heterogeneity is one-dimensional, and voters' preferences over a general income tax are monotonically related to their idiosyncratic productivity. The main result here is that the size of redistributive programs increases with a specific measure of pre-tax income inequality.

Section 3 studies the conflict between *young and old*. Now, there are two dimensions of heterogeneity, and voters' preferences over the generosity of the pension system are systematically related to their age, as well as their income. Large public pensions are supported by a coalition of poor and elderly voters, and the size of social security also exceeds the social optimum, because future generations of tax payers cannot participate in the voting.

Section 4 is devoted to the conflict between *employed and unemployed* individuals. It is then the employment status, or the risk of becoming unemployed, that shapes the preferences over the generosity of unemployment insurance and the structure of other labor-market programs. The powerful majority of "insiders" with stable jobs support an over-regulated labor market and under-provision of unemployment insurance.

Finally, Section 5 analyzes the conflict between *labor and capital*. Here, we study how the allocation of the tax burden between these two inputs is determined. Voters' preferences over the structure of the tax system predictably hinge on the relative importance of these two tax bases in their income. In equilibrium, taxes on capital are higher than what is socially optimal, since capital income is more concentrated and a majority of voters primarily rely on income from labor.

In Section 5, we also illustrate a different approach to representative democracy. This is the so-called *citizen-candidate* model, where elections are modeled as a contest between outcome-motivated candidates, who have explicitly chosen to undertake a costly entry decision in order to implement their ideologically preferred policy.

2. Rich vs. poor

How do voters evaluate redistributive programs? And how much income is redistributed? Can fundamental political forces account for the observed growth of social transfers over time, as well as the large cross-country differences in the size of these transfers, such as those mentioned in the General Introduction? These are questions motivating the literature surveyed in this section.

2.1. A simple model of redistribution

We start with a simplified version of a model originally proposed by Romer (1975) and Roberts (1977), and extended and popularized by Meltzer and Richard (1981). We reformulate the model slightly, to avoid unnecessary complications. Consider a static economy producing a single commodity. Individuals differ in one dimension only, namely their taxable income. As economic agents, they work and consume. As voters, they evaluate a simple redistributive program that pays a lump sum to each individual, financed by a *proportional* income tax. Below, we discuss how to introduce progressive taxation.

The preferences of the ith individual are:

$$w^i = c^i + V(x^i) ,$$

where c and x denote consumption and leisure respectively, and $V(\cdot)$ is a well-behaved concave utility function. The private budget constraint is

$$c^i \leqslant (1-\tau)\, l^i + f,$$

where τ is the income tax rate, l^i is individual labor supply, and f is a lump-sum transfer. The real wage is unity. Quasi-linear preferences imply that all income effects are absorbed by consumption. This simplifies the effect of tax distortions and the analysis of the voting equilibrium.

To model income differences, we assume that individual productivity differs, and that productivity, in turn, is equivalent to having more "effective time" available. That is, individuals are also subject to a "time constraint":

$$1 + e^i \geqslant x^i + l^i, \tag{2.1}$$

where e^i captures individual productivity. More productive individuals have a larger effective time endowment, e^i [5]. We assume that e^i is distributed in the population according to a known distribution with mean e, median, $e^m < e$, and a cumulative distribution function $F(\cdot)$.

It is easy to verify that in this simple model,

$$l^i = L(\tau) + (e^i - e), \tag{2.2}$$

where $L(\tau) \equiv 1 + e - V_x^{-1}(1-\tau)$ is decreasing in τ by concavity of $V(\cdot)$ [6]. Thus, as expected, a higher tax rate distorts the labor–leisure choice and induces the consumer to work less. By our assumption that $F(\cdot)$ is skewed, the distribution of income is skewed to the right, in conformity with available data in all countries.

Throughout this part, average variables are written without a superscript. Thus, l denotes average labor supply. Since the average of e^i is e, we have $l = L(\tau)$. The government budget constraint can therefore be written

$$f \leqslant \tau l \equiv \tau L(\tau). \tag{2.3}$$

Policy is set as follows. Two political candidates compete for office. They commit to electoral platforms formulated over the tax rate. Whoever wins the election enacts his

[5] The original model assumes that individuals only have different productivities when working, whereas we assume that more talented individuals are more productive at generating income as well as at enjoying their leisure time. As the next footnote shows, however, quasi-linear preferences imply that all individuals find it optimal to consume the same amount of leisure, while more talented individuals have more income and more consumption.

[6] Maximize the utility of individual i' subject to the budget and time constraints. The first-order condition implies: $1 - \tau = V_x(1 + e^i - l^i)$, where a subscript denotes a derivative. Take the inverse of $V_x(\cdot)$ and simplify to get the expression for l^i in the text. Note that $L_\tau = 1/V_{xx}(x^i) < 0$.

pre-announced policy. Both candidates are completely office-motivated, in the sense that they only care about winning the elections. They thus maximize the probability of winning[7].

2.2. Equilibrium redistribution

Consider the voters' preferences over policy. Define the indirect utility function of individual i, over τ, as:

$$W^i(\tau) \equiv \hat{c}^i + V(\hat{x}^i) \equiv (1-\tau)\hat{l}^i + \tau L(\tau) + V(1+e^i-\hat{l}^i), \tag{2.4}$$

where a $\hat{}$ refers to the private equilibrium choices, and where we used the private budget and time constraints and the government budget constraint to derive the right-most expression.

Let τ^i be the tax rate preferred by the ith individual. Then, τ^i is implicitly defined by the first-order condition $W^i_\tau(\tau^i) = 0$. We differentiate the right-most expression in Equation (2.4), noting that we can set $\mathrm{d}\hat{l}^i/\mathrm{d}\tau = 0$ by the envelope theorem. We then obtain:

$$W^i_\tau(\tau) = -\hat{l}^i + L(\tau) + \tau L_\tau(\tau) = -(e^i - e) + \tau L_\tau(\tau) = 0. \tag{2.5}$$

Consider the right-most expression of this condition. The first term is the marginal benefit of a higher tax rate cum redistribution. It is positive for a voter poorer than the average ($e^i - e < 0$) and negative for a voter richer than the average ($e^i - e > 0$). The last term is the marginal cost of higher distorting taxes, in the form of a smaller tax base; this term is always negative, as $L_\tau < 0$. Thus, each voter trades off the marginal redistributive benefit (or cost) of taxation against its deadweight loss. Equation (2.5) implicitly defines the tax rate preferred by voter i:

$$\tau^i = \frac{e^i - e}{L_\tau(\tau^i)}. \tag{2.6}$$

As $L_\tau < 0$, Equation (2.6) implies that a poor voter ($e^i < e$) prefers a positive tax rate, which is larger the poorer he is (the larger is e^i in absolute value), while a rich voter ($e^i > e$) prefers an income subsidy ($\tau < 0$), financed by a lump-sum tax. Individual preferences are thus monotonic in e^i. Furthermore, they are single-peaked by a natural restriction on $V(\cdot)$.

It is easy to see that there is only one political equilibrium: both candidates commit to τ^m, the policy preferred by the median voter. If any of the candidates were to announce a different value τ', the other candidate could ensure victory by announcing

[7] The argument is identical if we instead assume that candidates maximize their expected vote share.

a policy in the interval (τ', τ^m). Hence, the equilibrium tax rate, τ^m, coincides with the policy preferred by the median voter:

$$\tau^m = \frac{e^m - e}{L_\tau(\tau^i)}. \tag{2.7}$$

Up to the alternative citizen-candidate model presented in Section 5, all equilibria in this part can be thought of as the result of this kind of Downsian electoral competition.

2.3. Implications and evidence

The model thus predicts that the size of general redistributive programs reflects the preferences of the middle classes (the likely median voters), and is determined by their relative position on the income scale. By Equation (2.7), taxes are higher the greater the distance between median and mean income, a specific measure of income inequality. If the middle classes are *relatively* well off, because there is extreme poverty, equilibrium redistribution is small. If the middle classes are instead relatively worse off, with income highly concentrated at the top, equilibrium redistribution is large. Thus, the model predicts a link between skewedness of income distribution and the size of general redistribution schemes. Concentration of income at the top makes redistribution more attractive for the median voter, and hence increases the equilibrium tax rate. But more extreme poverty has the opposite effect, because it reduces the benefit of redistribution for the median voter. Another prediction of the model concerns the deadweight costs of taxation: the larger these are – as captured by the absolute value of L_τ – the smaller is equilibrium redistribution. Note, however, that the model really says nothing about selective, or targeted, transfer schemes, such as welfare payments.

Can this simple model explain secular growth in the size of redistributive programs and observed cross-country differences? Two features of the theory can possibly account for the early growth of redistribution. First, the extension of suffrage to poorer voters, early in the 20th century, certainly reduced the relative income of the median voter in Western democracies. Second, again early in the 20th century, economic progress and institutional change very likely reduced the transaction costs of collecting taxes, particularly income taxes, and hence the distortions associated with taxation. In the USA, for instance, income taxes only became constitutional in 1913. But what about the period after the 1960s? Electoral laws did not change and no major improvements in the tax-collection technologies occurred, and yet government transfers continued to increase as a fraction of national income?

Lindert (1994, 1996) systematically investigates these questions in a panel of OECD countries, in the periods 1880–1930 and 1962–1981, respectively. Running panel regressions that also control for average income, demographic structure of the population, and other variables, he finds conflicting results. On the one hand, voter turnout and redistributive transfers are positively related[8]. As voter participation

[8] See, for instance, Shields and Goidel (1997).

is positively correlated with relative income, this supports the theory. Moreover, high concentration of income (measured as the share of the top quintile relative to that of the middle quintile on the income scale) is indeed *positively* related to redistributive transfers, as predicted by the theory. Finally – though the evidence is somewhat weaker – poverty (the share of the bottom quintile relative to that of the middle quintile) is *negatively* related to government transfers, which is also a prediction of the theory. Income distribution can account for a large fraction of the observed cross-country differences in spending: the lower spending in the USA, in particular, could be attributed to lower voter turnout among poorer voters and to more extreme poverty, which raises the relative position of the median voter. On the other hand, when these measures of income distribution are replaced by the share of the middle quintile, which roughly measures the relative position of the median voter, it always turns out to be statistically insignificant.

The model we have discussed is static. Simple dynamic versions, where higher redistributive income taxation hurts the incentives to invest in physical or human capital and therefore economic growth, have been analyzed by Alesina and Rodrik (1994), Persson and Tabellini (1994a) and many others. More *wealth* inequality (in the sense of lower median relative to mean wealth) should thus be associated with higher taxation and slower growth. Alesina and Rodrik (1994) and Persson and Tabellini (1994a) find evidence in historical and cross-country data for more inequality indeed being associated with slower growth. But this is only indirect evidence, and the link between inequality and growth might be due to other economic or political mechanisms. Perotti (1996) indeed finds negative results when trying to relate various measures of income distribution to government transfers, in similar broad cross-country data. Data problems are, however, likely to be paramount in such broad data sets. The evidence from US states, where inequality data are more comparable, seems mixed[9].

Krusell and Rios-Rull (1999) instead focus on sequential voting decisions in a full-fledged dynamic economy. They calibrate a version of a neoclassical growth model with heterogeneity in wealth and labor income, where the same income tax applies to both labor and capital income. The model is formulated so that heterogeneity only affects political decisions, whereas only average magnitudes matter for the economic equilibrium. A median-voter result applies, similar to that illustrated above. But the median voter faces a more demanding problem: tax rates are chosen sequentially over time and the decision in each period is taken in full anticipation of how current policy influences the political equilibrium in the next period through its effect on the relevant state variables. Krusell and Rios-Rull numerically compute the political equilibrium and calibrate the steady state of the model to data for the US economy. Both steady-state tax rates and transfers are remarkably close to recent US data. Interestingly, the model's dynamics plays an important role: with fixed capital and variable labor supply

[9] See Partridge (1997) and Panizza (1999).

(a static version of the model) the same numerical calibration implies excessively high tax and transfer rates. Thus, the investment elasticity to the tax rate is important for quantitative success. Their paper, however, attempts to explain neither the secular rise of taxes and transfers, nor the observed cross-country differences.

Since a large component of tax revenues is used to finance non-redistributive public spending, the deadweight costs associated with the first dollar of redistributive spending can be very large (in the model they are zero, since all spending is assumed to be redistributive). Mulligan (2001) calibrates a static version of the model of this section for the USA and shows that this simple point can be quantitatively very important, and it can imply that only very poor individuals benefit from rich to poor redistribution. The US median voter, in particular, is unlikely to be poor enough to benefit much or at all from this kind of redistributive transfers.

Overall, these empirical results are disappointing: the secular increase in government transfers and the cross-country differences are huge, even if we restrict the sample to the last 30 years. A closer look at the timing of policy changes reveals a further weakness. In most countries, transfers rose most quickly in the 1960s and 1970s, when income inequality was generally on the decline; in the 1980s and 1990s, in contrast, inequality tended to increase once more, while redistributive transfers rose less quickly.

One reason why the theory may fail to account for the rise in government transfers in the last 30 years is that the data on transfers do not fit the theory very well. Pensions and health-related transfers are the most rapidly growing components of government transfers. As these systematically benefit older individuals, the simple median-voter model above needs to be modified to allow for heterogeneity in age. This is done in the next section. Other transfers belong to social insurance, such as transfers to the unemployed. This kind of spending also differs a great deal across countries. Section 4 investigates the determinants of unemployment insurance, which also differs from simple redistribution between rich and poor voters in several ways. Yet other transfers are very clearly targeted to more narrow groups. Such transfers, and the associated special-interest politics, are the topic of Part II.

2.4. Notes on the literature

The theory in this section is based on Romer (1975), Roberts (1977) and Meltzer and Richard (1981). It is straightforward to add public consumption, as an additional policy instrument or to replace the lump-sum transfer, provided the benefits of public consumption are not concentrated to particular income groups. Meltzer and Richard (1985) show that the same incentives to redistribute in cash then arise with respect to redistribution in kind. Cukierman and Meltzer (1991) replace the proportional income tax with a three-parameter tax schedule. Under plausible conditions on skewness of income distribution and labor-supply elasticity, a median-voter equilibrium exists and the decisive voter chooses marginal progressivity. Peltzman (1980) is an

influential early contribution, based on a very different political model and reaching very different conclusions.

A large empirical literature on the determinants of the size of redistributive programs is surveyed in Mueller (1989). The prediction that higher income inequality among voters leads to increased government redistribution has received particular attention in empirical studies. Lindert (1994, 1996) examines a panel of OECD countries with mixed results. The theory is instead supported by the analysis of US time-series data in Meltzer and Richard (1983), and by the different approach, based on calibration, of Krusell and Rios-Rull (1999), while the calibration exercise of Mulligan (2001) casts doubts on this simple model. Finally, Husted and Kenny (1997) show that the expansion of voting franchise is positively correlated with the size of redistributive programs by US states and local governments.

Empirical research also investigated the idea that the expansion of redistributive programs can be attributed to a reduction in the administrative costs of tax collection and the deadweight costs of taxation. Empirical support is provided by the works of Becker (1985), North (1985), Kau and Rubin (1981) and, more recently, Becker and Mulligan (1998).

A related literature has studied the links between redistributive policies, income inequality and growth. Alesina and Rodrik (1994) and Persson and Tabellini (1994a) provided the original impetus. Benabou (1996), Krusell, Quadrini and Rios-Rull (1997) and Persson and Tabellini (1999a) survey the theory on inequality and growth, while Perotti (1996) discusses the empirical findings on cross-country data. The evidence coming from US states, studied by Partridge (1997) and Panizza (1999), does not suggest strong conclusions. Arjona and Pearson (2001) find support for the link between inequality, redistributive transfers and growth in a panel of OECD countries between 1970 and 1998.

3. Young vs. old

Why have pension expenditures risen so rapidly in all countries in the postwar period, and with so little political opposition? What political forces stand in the way of pension reforms in most industrial countries? And how can a reform package be designed so as to be politically feasible? These are the questions motivating this section. We build on the simple median-voter model of Section 2, but add a second dimension of heterogeneity, age. As a result, public pensions redistribute both *across* and *within* generations. Intra-generational redistribution is a realistic feature of all pension systems and plays a key role in the political equilibrium, as voters' coalitions now form along two dimensions: age and income. The two-dimensional feature of coalition formation somewhat complicates the analysis. It is useful to study the complications in detail, however, as they illustrate how one may compute median-voter equilibria with multidimensional heterogeneity.

3.1. A simple model of pensions

Consider an overlapping-generations economy, where each generation lives for three periods and population growth is constant. There is no altruism across generations. Individuals work in the first two periods of life, and retire in the last period. They can invest their savings on a world-wide capital market at a given rate of return. Within each generation, labor income is heterogeneous. As in the previous section, some individuals have more effective time to allocate between labor and leisure; these productivity differences are permanent throughout life. A proportional income tax levied on working generations pays for the pensions of the retirees. A pension consists of the same non-negative lump-sum payment for every old individual. Thus, the pension system redistributes across and within generations. For simplicity, we treat the pension system in isolation from other parts of the budget; thus, taxes are only collected to finance pensions to the old under a balanced budget, whereas working generations receive no transfers.

When *young*, individual i maximizes the following utility function:

$$w^{iY} = U(c^{iY}) + \frac{1}{(1+\delta)}U(c^{iM}) + \frac{1}{(1+\delta)^2}c^{iO} + v(x^{iY}) + \frac{1}{(1+\delta)}v(x^{iM}), \tag{3.1}$$

where δ denotes the subjective discount rate; the notation otherwise coincides with that of the previous section, except that the upper-case superscripts denote the period of life. Linearity of consumption when old implies that all income effects are absorbed by c^{iO}. The intertemporal budget constraint of a young generation is

$$c^{iY} + \frac{c^{iM}}{1+\rho} + \frac{c^{iO}}{(1+\rho)^2} = l^{iY}(1-\tau) + \frac{l^{iM}(1-\tau)}{1+\rho} + \frac{f}{(1+\rho)^2}, \tag{3.2}$$

where ρ denotes the given world real interest rate, and f is the pension received when old. By assumption, the same tax rate τ is paid in both working periods (see further below). Finally, we assume that $\delta = \rho$. When choosing between labor and leisure, individuals face the time constraint (2.1) in each period, as in the previous section. This means that labor supply when young and (planned) labor supply when middle-aged are still given by Equation (2.2). Consumption when young and (planned consumption) when middle-aged are given by $c = U_c^{-1}(1)$, with income effects completely absorbed by consumption when old.

A *middle-aged* individual behaves in a similar fashion. He maximizes $(1+\delta)w^{iY}$, except that all variables from young age are now given. An *old* individual, finally, just consumes his pension plus his assets (or minus his liabilities).

Let n be the exogenous rate of population growth. Then, the government budget constraint can be written as:

$$f = \tau l^M(1+n) + \tau l^Y(1+n)^2 = \tau L(\tau)(1+n)(2+n), \tag{3.3}$$

where, as previously, non-superscripted variables denote averages. For each old individual, there are $(1+n)$ middle-aged and $(1+n)^2$ young individuals; the right-most expression follows from Equation (2.2) and some rewriting. This constraint (3.3)

is typical of a balanced pay-as-you-go pension system, where the contributions paid by the working generations finance the pensions of the currently old.

As real interest rates and real wages are both given in our simple perfect-foresight model, the pension system therefore has three economic effects only: it redistributes across generations, it redistributes within generations, and the taxes needed to finance it distort labor–leisure choices. In a richer model the pension system would have general-equilibrium effects via endogenous factor prices and would also provide social insurance in the face of individual income uncertainty.

3.1.1. *Voters' preferences*

How do different individuals evaluate the generosity of the pension system? Let us start with the simplest case of complete commitment to the system over time. Individuals are thus assumed to vote over τ (or, equivalently, over f). Once a policy is approved, it remains forever (or, equivalently, until all generations who voted for it have died).

All *old* voters clearly want the revenue-maximizing tax rate, as they only internalize benefits and no costs of higher taxes. Young and middle-aged individuals, however, base their policy preferences on both income and age. Generally, poorer and older individuals prefer higher public pensions, as they benefit more from either intra-generational or inter-generational redistribution.

Specifically, consider a young voter of type i, and let $W^{iY}(\tau)$ be his indirect utility function. By the envelope theorem, a marginal change in τ affects his welfare according to

$$\begin{aligned} W_{\tau}^{iY}(\tau) &= -\left[\hat{l}^{iY} + \frac{\hat{l}^{iM}}{1+\rho}\right] + \frac{1}{(1+\rho)^2}\frac{\mathrm{d}f}{\mathrm{d}\tau} \\ &= -\frac{(2+\rho)}{(1+\rho)}\left[L(\tau) + e^i - e\right] + \frac{(1+n)(2+n)}{(1+\rho)^2}\left[\tau L_{\tau}(\tau) + L(\tau)\right], \end{aligned} \tag{3.4}$$

where a $\hat{}$ denotes a privately optimal choice, as in the previous section, and where the right-most expression follows from Equations (2.2) and (3.3) and some manipulations. The expressions in Equation (3.4) are easily interpreted: increasing τ entails a benefit when old (the last term) and a cost in the first two periods of life, due to higher taxes (the first term). The benefit is the same for all young voters. But the cost of higher taxes is higher for the richer among the young (i.e., for those with a higher e^i). Moreover, higher population growth n makes public pensions more attractive, because the same tax rate now gives a higher pension. A higher real interest rate ρ would have the opposite effect, reducing the present value of net benefits from the pension system.

Consider the special, "golden-rule", case of $\rho = n$. Setting Equation (3.4) equal to zero, we get a condition identical to (2.5), that is, the condition for the optimal tax rate in the static model of the previous section! When $\rho = n$, the *average* young individual (with $e^{iY} = e^i = e$) gains nothing from the social-security system. But since taxes are distorting, he prefers $\tau = 0$. The social-security system becomes attractive for the

average young only if $\rho < n$. Young voters poorer than average ($e^i < e$), on the other hand, prefer $\tau > 0$ even if $\rho = n$, as they stand to gain from the *intra*-generational redistribution, just as in the model of the previous section.

Finally, consider a middle-aged voter of type i. By the same logic, a marginal change in τ affects his welfare according to

$$\begin{aligned} W_\tau^{iM}(\tau) &= -\hat{l}^{iM} + \frac{1}{1+\rho}\frac{\mathrm{d}f}{\mathrm{d}\tau} \\ &= -\left[L(\tau) + e^i - e\right] + \frac{(1+n)(2+n)}{(1+\rho)}\left[\tau L_\tau(\tau) + L(\tau)\right]. \end{aligned} \tag{3.5}$$

Comparing this expression with (3.4), the marginal benefit of pensions is now higher because it is closer in time, and the marginal cost is lower because taxes are now only paid for one period. Thus, a voter with the same relative income position $e^{iY} = e^{iM} = e^i$ prefers a higher tax rate when middle-aged than when young. In particular, the average ($e^{iM} = e$) middle-aged voter would prefer $\tau > 0$, even if $\rho = n$, though he would stop short of full revenue maximization.

By Equations (3.4) and (3.5), we can identify a pair of young and middle-aged individuals who always vote alike. Setting the right-most expressions in each of these conditions equal to zero, subtracting one from the other and simplifying, we get

$$e^{iM} = e^{iY} + \frac{(1+n)(2+n)}{2+\rho}\left[L(\tau) + \tau L_\tau(\tau)\right]. \tag{3.6}$$

For any young voter of type e^{iY}, there is thus always a middle-aged voter of type e^{iM} with identical policy preferences. This middle-aged voter is richer than his young counterpart, by Equation (3.6) $e^{iM} > e^{iY}$. The intuition was given above; older voters favor social security more as do poorer voters. Hence, for a young individual to prefer the same taxes as a middle-aged one, his lower age must be compensated by a lower income.

3.2. Equilibrium pensions

We are now ready to characterize the political equilibrium[10]. By the discussion above, individual preferences are single-peaked and monotonic in income and age. A median-voter result thus applies. But who is the pivotal voter? Clearly, all old individuals prefer the revenue-maximizing tax rate. Conversely, all young individuals richer than the average prefer tax rates at zero. The median voter will correspond to a pair: a poor young and a richer middle-aged voter, who prefer the same tax rate. Let e^{*m} be the *middle-aged* median voter (yet to be identified) – i.e., not the individual with the

[10] We only consider interior equilibria, such that $0 < \tau < \arg\max_\tau \tau L(\tau)$.

median endowment – and τ^{*m} his preferred policy. The relation between e^{*m} and τ^{*m} is obtained by setting the right-most expression in Equation (3.5) equal to zero, and solving for $e^i = e^{*m}$:

$$e^{*m} = e + \frac{(1+n)(2+n)}{1+\rho}\left[\tau^{*m}L_\tau(\tau^{*m}) + L(\tau^{*m})\right] - L(\tau^{*m}). \tag{3.7}$$

As before, let e^i be distributed in the population, with c.d.f. $F(\cdot)$. In equilibrium, the number of voters in favor of $\tau > \tau^{*m}$ equals the number of voters in favor of $\tau < \tau^{*m}$. By Equation (3.6), equilibrium requires that

$$\begin{aligned} &1 + (1+n)F(e^{*m}) \\ &+ (1+n)^2 F\left(e^{*m} - \frac{(1+n)(2+n)}{2+\rho}\left[L(\tau^{*m}) + \tau^{*m}L_\tau(\tau^{*m})\right]\right) \\ &\quad = \frac{1 + (1+n) + (1+n)^2}{2}. \end{aligned} \tag{3.8}$$

The left-hand side of Equation (3.8) is the size of the coalition of those voters in favor of taxes higher than τ^{*m}, namely all of the old and a fraction of the middle-aged and the young, respectively. In equilibrium, this coalition must make up exactly half the electorate, the measure of which is given by the expression on the right-hand side. We can also consider Equation (3.8) as an illustration of our previous claim: high pensions are supported by a coalition of *elderly and poor voters cum tax payers*, as those stand to benefit from the inter- or intra-generational redistribution.

To obtain the equilibrium policy τ^{*m}, combine Equations (3.7) and (3.8):

$$\begin{aligned} &(1+n)F\left(e + \frac{(1+n)(2+n)}{1+\rho}\left[\tau^{*m}L_\tau(\tau^{*m}) + L(\tau^{*m})\right] - L(\tau^{*m})\right) \\ &\quad + (1+n)^2 F\left(e + \frac{(1+n)(2+n)}{(1+\rho)(2+\rho)}\left[\tau^{*m}L_\tau(\tau^{*m}) + L(\tau^{*m})\right] - L(\tau^{*m})\right) \\ &= \frac{(1+n) + (1+n)^2 - 1}{2}. \end{aligned} \tag{3.9}$$

As F is a monotonic function, Equation (3.9) implicitly defines a unique equilibrium tax rate. This tax rate τ^m is a decreasing function of ρ: a higher ρ reduces the present discounted value of future pensions, making young and middle-aged voters less favorable to public pensions. A higher population growth rate n, on the other hand, has ambiguous effects on τ^m. On the one hand, a higher n increases the weight of the young and reduces the weight of the old, thus shifting the median-voter identity towards someone less favorable to pensions. On the other hand, a higher n makes pensions more attractive for all young and middle-aged voters. Either effect might prevail, depending on functional forms. Finally, the shape of the income distribution, as described by $F(\cdot)$, also affects equilibrium policy. But then, not only median income matters, as the

decisive voters are not median-income recipients. In general, more income inequality is likely to make the decisive voters more willing to exploit the pension system for intragenerational redistribution, and increase the equilibrium generosity of the system.

It is useful to consider the special case where $\rho = n$. Here, it can be shown that τ^m is larger than the equilibrium tax rate of the static redistributive model in the previous section. In fact, the two tax rates would coincide if only the young individuals were eligible to vote. As noted above, if $\rho = n$ the young do not benefit from the intergenerational redistribution, and only the intra-generational motives for redistribution would shape their votes. But the old and middle-aged do benefit from intergenerational redistribution, even if $\rho = n$. Their votes thus raise the equilibrium generosity of the pension system beyond what the median young individual prefers.

Suppose we let the normative benchmark be a utilitarian optimum, defined as the maximal discounted sum of the welfare of all currently alive and future generations. By the quasi-linearity of preferences, this translates into a discounted sum of the welfare of the average individual in each generation. It is easy to see that the utilitarian optimum has $f = 0$ [11]. Relative to this benchmark, the political equilibrium we have studied entails too much redistribution, both across and within generations. First, it redistributes to poor individuals at the expense of rich. As in the previous section, this is a consequence of majority rule and the distribution of income being skewed to the right. Second, the equilibrium redistributes to the currently old and middle-aged voters, at the expense of future generations. This new feature is a consequence of the yet unborn generations not participating in the vote determining their future taxes. There are thus powerful political forces supporting the introduction of a pay-as-you-go, social-security system, and keeping its size excessive relative to the social optimum.

One of the political distortions that keeps public pensions too large is that future generations are affected by the system, but do not vote on it. This suggests a simple constitutional remedy: only the young generation should be allowed to vote on social security, since it is the only generation that correctly internalizes the entire tax burden of public pensions. Naturally, this constitutional constraint is hard to enforce, as there would always be a majority of voters willing to repeal it. Moreover, the political equilibrium described above hinges on the assumption of commitment; once voted upon, the policy remains for as long as all generations participating in the vote are alive. Below, we discuss how to relax this assumption.

3.3. Evidence and extensions

We have just illustrated how political forces may bring about and shape a pension system of the kind observed in many western democracies. Does the evidence support

[11] There may be other reasons, such as social insurance, for positive socially optimal pensions. As these are not included in our simple model with risk neutrality, the argument should be interpreted as deviations from some benchmark, whatever the level of pensions in that benchmark.

some of the specific predictions of the model? There are few empirical studies. The demographic composition of the population is clearly an important determinant of the size of pensions. Lindert (1996), Perotti (1996) and Tabellini (2000) all find that, in panels of industrial countries and in cross-sectional correlations of larger country groups, pension expenditures *as a fraction of GDP* are larger the greater is the share of elderly in the population. But this finding does not discriminate well against other possible models of equilibrium pensions. A social planner would also spend more on pensions, if there were a larger number of elderly. The model's prediction is really that pensions *per retiree* would be higher, the higher the weight on old voters (a lower n in the model), as this shifts the median-voter equilibrium towards a more generous pension system[12]. Further, population growth is, in reality, not constant over time. Being faithful to the theory, one should also look at the effect of changes in expected future population trends (this is the second and opposite effect of n on the political equilibrium above). But no empirical study of which we know incorporates these features, nor has anyone studied the effect of the real interest rate, ρ.

The model also predicts pensions to increase with appropriate measures of income inequality. This is only very weakly supported by the evidence. Lindert (1996) and Perotti (1996) find no significant effect of income distribution variables on pensions. Tabellini (2000), on the other hand, finds a positive correlation between a Gini index of inequality and pensions in a large sample of countries, controlling for age and initial income. But measures of inequality are bound to be highly imperfect for such a large sample of countries. And measuring income distribution in accordance with the model is even more tricky; as noted above, the relative income of the decisive voter is age-dependent and does not coincide with median income.

The prediction that individual policy preferences over public pensions depend on age and relative income can also be tested directly, independently of the equilibrium predictions of the median-voter model. This has been done by Boeri, Börsch-Supan and Tabellini (2001), by means of opinion polls. In a recent survey of the opinions of 5000 European citizens, they find that individual willingness to opt out of the pay-as-you-go pension system is systematically related to age and income, with younger and richer individuals more willing to opt out, as expected. In the same survey, they also find two other relevant empirical results. On the one hand, a majority of the respondents in each of the four countries studied (France, Germany, Italy and Spain) opposes reforms that would shrink the size of the welfare state. This is exactly what the median-voter model predicts about the status quo (contrary to the predictions of models of lobbying and probabilistic

[12] In a cross-section study of social spending in Swedish municipalities, Strömberg (1996) explicitly tests – and finds support for – a political model based on the age of the median voter against a social-planner alternative.

voting discussed in Part II). On the other hand, the survey also reveals a huge ignorance of the true cost of the pay-as-you-go pension system, shedding some doubts on the assumption of voter's rationality that figures prominently in all of political economics.

The simple model studied in this section has been generalized in many directions. If we add capital accumulation, the social-security system generates *general-equilibrium* effects, at least in a closed economy. An expansion of the program reduces private savings, raises the real interest rate and lowers the real wage. This benefits rich savers and hurts borrowers, thus adding another dimension to the political determinants of the equilibrium. As Cooley and Soares (1999) show, these general-equilibrium effects can sometimes play a dominant role in studying the preferences over the pension system.

With individual income uncertainty, the pension system also has *social-insurance* benefits. Conesa and Krueger (1999) incorporate both types of effects in their analysis of the political support for pension reform. They study a rich model with heterogeneity in three dimensions: age, assets and income. Conesa and Krueger use numerical methods to study the economy's dynamic adjustment over time to different types of pension reform. Their results illustrate clearly how hard it is to muster majority support among the present voters for reforms of the pension system, even though the reforms bring about significant long-run benefits.

The assumption of *commitment* can also be modified without altering the nature of the results. Suppose that the effect of majority decisions only lasts one period, rather than forever, as assumed above. Thus, every other period, voters get to decide on social-security contributions today and tomorrow. In the absence of reputational effects, all young voters would now vote against any positive contributions, since the pension they will receive two periods hence is not affected by the current vote. The old and a fraction of the middle-aged individuals, on the other hand, continue to support the social-security system. Unless they are outnumbered by the young, the same factors as above, namely ρ, n and the function $F(\cdot)$ will shape the equilibrium policy, even though the precise characterization will differ and the system will be less generous. Indeed, such a model of limited commitment might be a good vehicle for studying the evolution of social security over time, in the face of changing population trends. The ongoing and predicted aging of the population, experienced in most western democracies, would introduce interesting dynamics in public support for the pension system. Studying these dynamics might give a deeper understanding of the forces behind the political struggle over pension reform.

Absent any commitment, positive pensions could not be sustained in the simple model of this section, except through reputational forces. All taxpayers would oppose the system, as their pension would be independent of the current vote. Presumably, two generations of taxpayers would also outnumber the old generation. No commitment is, however, as unrealistic as full commitment. Abolishing the pension system from one day to the next would not only meet political resistance not fully captured by our simple median-voter model, but would be ruled out as unconstitutional in many countries. Moreover, reputational mechanisms could link the voting outcomes across

period, and in this case equilibria with social security could even be sustained without commitment[13].

Altruism across generations is another mechanism that may help sustain equilibria with positive social security in the absence of commitment. Tabellini (1991, 2000) shows that, even if altruism is so weak that it will not support private inter-vivos transfers, it can nevertheless affect political behavior. Poor young and middle-aged individuals could be induced to vote in favor of the social-security system, because the (lump-sum) benefit to their parents or grand-parents outweighs the cost of the small taxes they must pay. With a larger number of poor voters than rich, this could be enough to support public pensions.

An important assumption in our model is the restriction to just one policy instrument: a lump-sum transfer when old financed by a wage tax. Yet, as emphasized by Mulligan and Sala-i-Martin (1999a), social-security systems are characterized by a variety of policy dimensions, and a good theory ought to be able to explain most of them. In this vein, Galasso and Conde Ruiz (1999) consider an overlapping-generations model combining the two redistributive policy tools: a purely *intra*generational scheme of redistribution, like that of Subsection 2, and the pension scheme of the present subsection that redistributes both within and across generations. In their model, preferences are no longer single-peaked, and they study a structure-induced equilibrium as in Shepsle (1979). They find that both redistributive tools are used in equilibrium. But, consistent with the evidence, the intragenerational scheme turns out to be much smaller than the pension system. The reason is that the old are a homogeneous and large coalition, who supports pensions but not other forms of redistribution. This may also help explain why pensions are financed out of wages, with no explicit or implicit taxes on accumulated wealth (pension benefits are almost never conditional on individual wealth holdings): taxing wealth would break the homogeneity of the old generation vis a vis the policy, and reduce the size of the coalition in favor of larger pensions. Another important policy dimension is the age of retirement. Why is retirement compulsory in virtually all public-pension systems? And what determines retirement age? As suggested by Mulligan and Sala-i-Martin (1999b), compulsory retirement is likely to increase the political influence of the elderly: not having other sources of income makes the economic interest of the old more homogeneous and increases their stakes. This by itself is likely to increase their political influence, as the models of probabilistic voting or lobbying of Part II would imply. Mulligan and Sala-i-Martin (1999b) focus on yet another aspect: retirement frees up leisure time, that can be devoted to lobbying and other political activities.

[13] Reputational equilibria in overlapping-generations (OLG) models may be quite different from the usual applications of the folk theorem, in that they may require generational chains of punishments or rewards. In a simple two-period OLG model, for example, sustaining a reward from the current young to the current old requires that the current young expect that their hypothetical deviation from the equilibrium would lead to future punishment from the next (yet unborn) generation.

3.4. Notes on the literature

The theory of voting over social security has followed different approaches. Browning (1975) and Boadway and Wildasin (1989a,b) study the determinants of social security in voting models with commitment, where all voters have the same income and differ only in age. Cukierman and Meltzer (1989) consider public debt (equivalent to social security in their model) in an overlapping-generations economy with income heterogeneity, weak altruism within the family, and policy commitments. Tabellini (2000) formulates a median-voter model with income heterogeneity and weak altruism within the family, but no commitment (i.e., in each period, voters choose a tax rate with lump-sum transfers to the currently old). The model of the present section, where voters differ in age and income, but where there is commitment and no altruism, combines features of all these approaches. These results are perhaps closest to those of Cukierman and Meltzer (1989), though that paper focuses on general-equilibrium effects on the real interest rate and neglects tax distortions. General-equilibrium effects and their effect on voters' preferences have also been studied by Cooley and Soares (1999). Conesa and Krueger (1999) include in their analysis not only general-equilibrium effects, but also social-insurance benefits of the pension system. A general survey of the positive political theories of social security is provided by Verbon (1988), while Feldstein (1998) and Siebert (1998) discuss the recent reform experiences of various developing and industrial countries.

In the absence of policy commitment, social-security systems can be sustained by reputational equilibria. This idea was pursued by Kotlikoff, Persson and Svensson (1998), and more recently by Boldrin and Rustichini (2000), Cooley and Soares (1999) and Azariadis and Galasso (1997). The idea that altruism within the family also induces voters to support intergenerational redistribution is investigated by Tabellini (1991, 2000).

Some papers have studied the political determinants of social security in settings different from voting. Grossman and Helpman (1998) consider a model where members of different generations lobby the government, as in Part II below. Earlier papers relying on the idea that the ability of different generations to influence the political process affects the size and viability of social security include Patton (1978), Stuart and Hansson (1989) and Loewy (1988). More recently, Lambertini and Azariadis (1998) have focused on legislative bargaining among (representatives of) different interest groups. Mulligan and Sala-i-Martin (1999b) and Galasso and Conde Ruiz (1999) study multidimensional aspects of pension policy in slightly different political models.

The validity of the empirical prediction that more inequality leads to more spending on social security has been investigated by Lindert (1994, 1996) with negative results, whereas Tabellini (2000) obtained more encouraging results. Looking at data of Swedish municipalities, Strömberg (1996) finds support for the prediction that the composition of social spending is systematically related to the age of the median voter.

The opinions of European citizens towards welfare-state programs and pension systems have been studied in Boeri, Börsch-Supan and Tabellini (2001).

4. Employed vs. unemployed

In the previous sections, voters knew their relative income with certainty when choosing their policy. An important role of some redistributive transfer programs, however, is to provide insurance against income risks, as in the case of unemployment insurance or public health insurance[14]. Voters evaluate such programs on the basis of their relative risk, besides their relative income. In labor markets, the distribution of risk among individuals is also affected by government regulation, such as hiring and firing rules. This section analyzes the political determinants of unemployment insurance as well as labor-market regulations.

A central determinant of such programs, emphasized by Wright (1986) and Saint-Paul (1993, 1996), is the likely conflict of interest between employed and unemployed voters, or more generally between *insiders* (those with a well-paid and protected job) and *outsiders* (the unemployed and workers in secondary markets). To keep things simple, we abstract from idiosyncratic unemployment risk, even though risk differences are realistic and could be added. The remaining conflict of interest then becomes very stark: the risk of future unemployment is lower for currently employed workers/voters, who want less unemployment insurance than the unemployed. Instead, currently employed voters find it more expedient to protect themselves against unemployment risk through tight firing restrictions, even though such restrictions would increase unemployment and unemployment duration. As employed voters constitute a majority, political equilibria generally exhibit underprovision of unemployment insurance and overly restrictive labor-market regulations. Closing the section, we discuss how labor-market reforms may become politically feasible.

4.1. *A simple model of unemployment insurance*

All individuals are alike, apart from their employment status, and they maximize expected discounted lifetime utility of consumption over an infinite horizon:

$$V^J = \mathsf{E}_0\left[\sum_{t=0}^{\infty}\beta^t U(c_t^I) \mid I = J \text{ at } t = 0\right], \quad I,J \in \{E, U\},$$

where E_0 is the expectations operator conditional on information available at time 0, t is the time period, β is a discount factor [$\beta = 1/(1+\delta)$ in the notation of Section 3],

[14] We rarely observe private unemployment insurance. But we do not discuss the underlying informational problems, which presumably provide a rationale for government insurance. It is not straightforward, however, to provide such a rationale. Under moral hazard, a government facing the same information constraints as private agents would not generally be able to outperform the market. Under adverse selection, there is more scope for outperforming the market, as the government might rely on compulsion.

and $U(\cdot)$ is a well-behaved concave utility function. Individuals are either employed or unemployed and the E and U superscripts denote these two states. Labor supply is exogenous and set equal to one. For simplicity, we also assume that there are no credit markets (see further below). Hence, unemployment insurance entails no distortions, and consumption equals current income. If employed, individuals thus consume their real wage, normalized to unity, less taxes, $c_t^E = 1 - \tau_t$. If unemployed, they receive an unemployment benefit, c_t^U.

Individual employment status follows an exogenous stochastic (Markov) process. In each period, a currently employed individual becomes unemployed with probability φ (for firing rate), whereas a currently unemployed individual becomes employed with probability ϑ (for hiring rate). By the Markov assumption these transition probabilities remain constant over time, irrespective of an individual's employment history, and are the same across individuals. The aggregate rate of unemployment u_t is given by

$$u_t = \varphi(1 - u_{t-1}) + (1 - \vartheta)\, u_{t-1}. \tag{4.1}$$

In each period, unemployment consists of the previously employed who were laid off (the first term), plus the previously unemployed who did not find a job (the second term). We focus on the steady state, where u_t has converged to a constant. Solving Equation (4.1) for $u_t = u_{t-1} = u$ yields:

$$u = \frac{\varphi}{\varphi + \vartheta}. \tag{4.2}$$

We assume that $\varphi + \vartheta < 1$ and that $\vartheta > \varphi$, so that u is less than 50%.

Finally, as in the previous sections, we treat this government program in isolation from other policies. The government budget constraint implies that unemployment subsidies must be financed by taxes on currently working individuals:

$$uc_t^U = \tau_t(1 - u).$$

Using Equation (4.2), the government budget constraint can be written as:

$$c_t^U = \tau_t \frac{\vartheta}{\varphi}. \tag{4.3}$$

4.1.1. Voters' preferences

Assume initially that unemployment insurance is chosen today (at $t = 0$), *given* that u is already at its steady-state value, and stays in place forever: that is, $\tau_t = \tau$, and $c_t^U = c^U$ for all t. How do voters evaluate such a program? To answer this question,

consider the value functions of employed and unemployed voters, respectively. Making use of the previous expressions for c^E and c^U, these can be written as

$$\begin{aligned} V^E &= U(1-\tau) + \beta[(1-\varphi)V^E + \varphi V^U], \\ V^U &= U(\tau\frac{\vartheta}{\varphi}) + \beta[\vartheta V^E + (1-\vartheta)V^U]. \end{aligned} \tag{4.4}$$

The solution yields the state utilities as a function of the policy τ:

$$\begin{aligned} V^E &= \frac{\beta\varphi U(\tau\frac{\vartheta}{\varphi}) + (1-\beta(1-\vartheta))U(1-\tau)}{(1-\beta)(1-\beta(1-\vartheta-\varphi))}, \\ V^U &= \frac{(1-\beta(1-\varphi))U(\tau\vartheta/\varphi) + \beta\vartheta U(1-\tau)}{(1-\beta)(1-\beta(1-\vartheta-\varphi))}. \end{aligned} \tag{4.5}$$

Taking the derivative of these expressions with regard to the policy τ, and setting it equal to zero, we find the insurance policy desired by employed and unemployed individuals, respectively:

$$\begin{aligned} \frac{U_c(c^E)}{U_c(c^U)} &= \frac{\beta\vartheta}{1-\beta(1-\vartheta)} \leqslant 1, \\ \frac{U_c(c^E)}{U_c(c^U)} &= \frac{1-\beta(1-\varphi)}{\beta\varphi} \geqslant 1, \end{aligned} \tag{4.6}$$

where the inequalities follow from $\beta \leqslant 1$. Evidently, the currently employed prefer incomplete insurance ($c^E \geqslant c^U$), while the currently unemployed prefer over-insurance ($c^U \geqslant c^E$). Even though both sets of voters face a probability of changing status, the dominant force is still that current unemployment insurance redistributes from employed to unemployed voters (unless $\beta = 1$). By contrast, a utilitarian social planner – equivalently, an individual who maximized his expected utility behind a veil of ignorance over his current employment status – would always prefer full insurance, $c^E = c^U$. This is intuitive, as there is neither aggregate risk, nor individual incentive problems due to information or distortive taxation. Adding such inefficiencies would lower the desired insurance levels discussed above, but not eliminate the conflict between employed and unemployed.

Note that the qualitative results do not hinge on the absence of credit markets. With perfect credit markets and no aggregate risk, individuals would be able to fully insure their unemployment risk. Yet, some individuals would still want to use public unemployment insurance to redistribute in their favor. In particular, unemployed voters, or more generally voters whose risk of being unemployed is higher than average, would want public unemployment insurance since it would redistribute towards them in expected value [discussed by Wright (1986)]. If private insurance markets were absent but individuals could still save, they would have an incentive to self-insure.

As a result, the tax cost of financing the program would weigh more heavily in their preferences and they would prefer less unemployment insurance in the labor market, a point emphasized by Hassler and Rodriguez Mora (1999).

4.2. Equilibrium unemployment insurance

With only two types of voters, the political equilibrium is simply the policy preferred by the largest group, namely those currently employed[15]. To get explicit results, let the utility function be iso-elastic, $U(c) \equiv c^{1-\gamma}/(1-\gamma)$, with γ denoting the coefficient of relative risk aversion. The first expression in Equation (4.6), and the definitions of c^E and c^U, imply that the equilibrium tax rate τ^E satisfies

$$\frac{\tau^E \vartheta}{(1-\tau^E)\varphi} = \left[\frac{\beta\vartheta}{1-\beta(1-\vartheta)}\right]^{1/\gamma}. \tag{4.7}$$

From the government budget constraint (4.3), we can easily derive the corresponding equilibrium unemployment benefit, c^U.

How is equilibrium policy affected by changes in the parameters of the model? The implicit function theorem implies

$$\begin{aligned} &\frac{\partial\tau^E}{\partial\varphi} > 0, \quad \frac{\partial c^U}{\partial\varphi} < 0; \qquad \frac{\partial\tau^E}{\partial\vartheta} \lesseqgtr 0, \quad \frac{\partial c^U}{\partial\vartheta} > 0; \\ &\frac{\partial\tau^E}{\partial\beta} > 0, \quad \frac{\partial c^U}{\partial\beta} > 0; \qquad \frac{\partial\tau^E}{\partial\gamma} > 0, \quad \frac{\partial c^U}{\partial\gamma} > 0. \end{aligned} \tag{4.8}$$

A higher firing rate φ reduces the equilibrium unemployment benefit but raises the equilibrium tax. Intuitively, with a higher firing rate, employed voters still want to retain the same marginal rate of substitution between consumption if employed or unemployed, as is evident from Equation (4.6). But that rate has become more expensive, as equilibrium unemployment is larger, as is evident from Equation (4.2). It is optimal to adjust both margins, raising the tax rate but reducing the unemployment benefit. Conversely, if the hiring rate ϑ is higher, the risk of becoming unemployed is less menacing, and the decisive voter is willing to accept a higher marginal rate of substitution of consumption if employed vs. unemployed. In this sense, less insurance is needed. But insurance is now cheaper to buy, because unemployment falls with a higher ϑ. Hence, the unemployment benefit rises and the tax rate falls,

[15] The equilibrium generalizes to the case discussed above of idiosyncratic unemployment risk, when the latter is modeled as idiosyncratic hiring and firing parameters ϑ^i and φ^i. The political equilibrium would still be a median-voter equlibrium even with such two-dimensional heterogeneity. But as in the case of pensions, the decisive voter would be a pair, namely an employed high-risk type and an unemployed low-risk type with different values for φ^i and ϑ^i.

only if the individual is sufficiently risk averse (more precisely, if $\gamma > 1$)[16]. A higher discount factor or a higher rate of risk aversion, finally, would imply a more generous program, as the future risk of unemployment now carries more weight in the decision.

4.2.1. *Evidence and extensions*

From a positive point of view, it is interesting to note that the unemployment benefit, c^U, is negatively related to unemployment: parameter changes which increase unemployment also reduce the unemployment benefit. The reason is that the decisive voter reacts to changes in the cost of providing unemployment insurance[17]. The model also has unambiguous predictions regarding the effect of the general turnover in the labor market on the generosity of equilibrium unemployment insurance. To see this, consider a fall in both ϑ and φ such that the ratio ϑ/φ, and hence aggregate unemployment u, stay constant. It is easily shown that both τ and c^U decrease for such an increase in turnover. It is unclear whether these two predictions are consistent with the development over time of unemployment insurance in European countries, where indeed unemployment has generally increased and turnover in the labor market has generally decreased over the last two decades. It has not yet been explored whether these predictions are consistent with the evidence, even though it would seem feasible and well worth the effort. It is quite clear, however, that the model's predictions for Europe vs. the USA are counterfactual: Europe has both higher unemployment and lower turnover, at least in recent times, but higher unemployment benefits.

Such counterfactual cross-sectional predictions motivated Hassler and Rodriguez Mora (1999) to study the role of self-insurance. They show that once self-insurance is allowed, higher turnover does indeed make the employed prefer less generous unemployment insurance: when turnover is high, private savings become a close substitute for unemployment insurance, making the latter less valuable. Hassler and Mora also discuss the difficulty of sustaining positive unemployment insurance if there is no commitment to policy in future periods; this point is closely related to our discussion about the sustainability of the pension system in the previous section.

Our simple model of *endogenous* policy above focuses on the link from unemployment and its determinants to unemployment benefits. Much of the traditional literature on *exogenous* policy discusses the link in the opposite direction. That is, generous unemployment benefits may generate higher unemployment, either by pushing up

[16] Note that the above are pure comparative-statics experiments. Specifically, they assume that a parameter difference has fully manifested itself in a different steady-state unemployment rate before the choice of unemployment insurance takes place.

[17] These comparative-statics results would be less clear cut with individual specific hiring and firing rates. In that case, parameter changes would alter the identity of the median voter and, as unemployment increases, the median voter would be more likely to be unemployed. This would tend to move the size of equilibrium unemployment insurance (also as measured by benefits) in the same direction as the rate of unemployment.

equilibrium wages or by pushing down equilibrium search effort[18]. In an interesting paper, Hassler et al. (1998) try to incorporate both links in a model with labor-market search and endogenous policy. They show that there may very well be multiple equilibria: one with high unemployment and generous benefits and another with low unemployment and less generous benefits.

4.3. Equilibrium labor-market regulations

Unemployment insurance is not the only policy where the preferences of the employed and the unemployed clash. Labor markets in many industrial countries, particularly in Europe, are heavily regulated. In particular, firings are restricted or costly for the firm, not by contract, but by law. These regulations protect those currently employed but harm the unemployed, since they discourage new hires and thus increase unemployment duration. We now investigate the political determinants of these regulations, largely following Saint-Paul (1996).

Consider the same economy as above, but without public unemployment insurance: the unemployed earn a given subsistence wage, and consumption of the employed is exogenously given[19]. To model firing regulations, redefine the probability of becoming unemployed, φ, as:

$$\varphi = \chi + q,$$

where q is voluntary quits, and χ is firing (lay-offs) by the firms. We treat q as an exogenous parameter, but χ as a policy variable. The latter captures the influence on firings of specific labor-market legislation. The more difficult it is to legally fire a worker, the lower is χ and, hence, the lower is φ. We can thus interpret χ as a measure of labor-market flexibility: a higher χ amounts to more flexibility. As discussed by Saint-Paul (1996), who uses earlier results by Pissarides (1990), firing restrictions also make firms less willing to post vacancies. Thus, firing restrictions reduce the hiring rate, ϑ. Specifically, suppose – as does Saint-Paul (1996) – that the hiring rate is a given concave function of the firing rules:

$$\vartheta = H(\chi) \quad \text{such that} \quad H_\chi > 0,\ H_{\chi\chi} < 0. \tag{4.9}$$

That is, more flexible labor markets allow firms to increase firings (χ increases) but also tend to increases the hiring rate, though at a decreasing rate. Firms are thus assumed

[18] Layard and Nickell (1999) survey the relevant literature.

[19] This rules out general-equilibrium effects of changes in the unemployment rate, operating through the government budget constraint. These effects would make the voting problem dynamic, as voters would have to consider the dynamic adjustment to the steady state – recall that by Equation (4.1), unemployment gradually adjusts to the steady state. While these dynamic effects are unlikely to overturn the conclusions of this subsection, they complicate the analysis considerably.

to be more willing to hire workers, if they know it is easier to lay them off during bad times. This means that increasing labor-market flexibility involves a trade-off between firing and hiring rates. This trade-off is more favorable when labor markets are very rigid, that is when χ is low, for the hiring rate increases more, in this case, as a result of increased flexibility.

This formulation implies that labor-market flexibility generally has an ambiguous effect on steady-state unemployment, depending on the value of χ. In fact, by Equation (4.2):

$$\frac{\partial u}{\partial \chi} = \frac{H(\chi) - (\chi + q) H_\chi(\chi)}{(\varphi + \vartheta)^2} \lesseqgtr 0. \tag{4.10}$$

By concavity of $H(\chi)$, this derivative is more likely to be negative for low values of χ. That is, additional labor-market flexibility is more likely to reduce unemployment when labor markets are very rigid, due to the greater marginal effect on hiring. We make this explicit by assuming that $u(\chi)$ – i.e., unemployment as a function of labor-market flexibility for given q – has a unique minimum $u(\widetilde{\chi})$ at a specific level of labor flexibility $\widetilde{\chi}$.

This simple model is obviously a short-cut, in that it does not treat firm behavior explicitly, squeezing what is essentially a dynamic problem into a static reduced-form hiring function. The ambiguous effect of firing protection on unemployment, due to the opposite reaction of the firing and hiring rate, is a well-known property also of more sophisticated theoretical models of unemployment; the ambiguity is often the basis of arguments that easier firing rules would not necessarily help reduce the high European unemployment – see, for instance, Mortensen and Pissarides (1999) for a survey of the theoretical literature on the natural rate of unemployment, and Blanchard and Wolfers (2000) on European unemployment.

Without further excuses, we now turn to the political equilibrium. Clearly, employed and unemployed voters disagree over flexibility: the currently employed insiders want to protect their jobs, and thus dislike flexibility, while the unemployed outsiders welcome flexibility as it raises the hiring rate. The unemployed constitute a minority, however, and equilibrium policy is thus chosen to please the employed voters.

Formally, the equilibrium policy is the value of χ which maximizes the employed voters' expected lifetime utility. As in the previous subsection, the maximand is given by V^E in Equation (4.4), except that φ is now replaced by $\chi + q$ everywhere. The first-order condition for χ is obtained by taking the partial derivative of V^E with respect to χ, given (4.9), and setting it equal to zero. After some rewriting, we can express the equilibrium condition as

$$H(\chi) - (\chi + q) H_\chi(\chi) = -\frac{1-\beta}{\beta}. \tag{4.11}$$

The right-hand side of Equation (4.11) is strictly negative (as $\beta < 1$). But then it follows from Equation (4.10) that, in equilibrium, $\partial u/\partial \chi < 0$. That is,

equilibrium unemployment is above its minimum, defined by $u(\tilde{\chi})$, and would be reduced by additional labor-market flexibility. To protect their jobs, the majority of employed voters restricts firing to the extent that unemployment increases. This also has costs for the insiders, however. If unemployed in the future, they will have to wait longer for a job. At some point, these costs of unemployment become high enough to outweigh the benefits to insiders of tighter labor-market restrictions[20].

This result, that high equilibrium unemployment is also caused by overly tight firing rules, contrasts with the previously quoted arguments, i.e., increasing labor-market flexibility would not necessarily reduce European unemployment. These arguments are based on an incomplete theory, however, as they view the level of existing regulations as random. But policy choices are certainly not random: existing labor-market regulations largely reflect the preferences of the majority of "insiders". If so, their predicted effect on unemployment is clear: easier firing rules, if politically feasible, would reduce unemployment. The view that existing policy choices are not random, but systematically related to the political and economic environment, also has important implications for how to approach empirically the unemployment effects of alternative labor-market policies and institutions. These implications have, so far, been neglected in the existing empirical literature on the economic causes of unemployment – see Layard and Nickell (1999) for a very good survey.

4.3.1. Extensions

Are there policy reforms that retain job security for insiders and, at the same time, reduce unemployment? If so, they would clearly be politically feasible, for they would receive the support of both employed and unemployed voters. Higher public employment could be one solution. Marginal employment subsidies or other devices to stimulate labor demand by private firms would be another solution. In both cases, however, some taxpayers would have to foot the bill. It would also be more difficult to fully analyze the equilibrium provision of alternative public policies. One way would be to combine this model with the one studied in Section 2, where there is income heterogeneity among employed workers and the tax burden is not evenly shared among these.

Saint-Paul (1996) discusses other paths to reform. One is labor-market segmentation. Suppose the law would provide two kinds of firing restrictions: tighter ones for old jobs, but looser restrictions (or no restrictions at all) for new jobs. Such a two-tier system would protect the job security of insiders, while, at the same time, reducing unemployment. Thus, it would be an improvement for all voters, and would be supported politically. In the long run, a problem might emerge, however. As more

[20] With endogenous income taxes or unemployment subsidies, there would be a further cost of higher unemployment: providing unemployment insurance becomes more expensive, as taxes must increase or, equivalently, lower unemployment benefits can be financed out of given tax revenues.

and more workers would become employed on more flexible contracts, insiders might become a political minority in the sense that their labor-market protection could be scrapped and their rents eroded. Expectations of this long-run outcome could reduce the support of insiders for a two-tier labor market. Saint-Paul (1996) shows a possible solution. Less protected jobs should only remain so temporarily. That is, the law should specify a conversion clause: after some time, new jobs should either become regular and enjoy the full benefits of tight firing rules, or they should be scrapped. Such a reform would still reduce unemployment, without adverse long-run political consequences.

Research on these issues is still scarce. High equilibrium unemployment has become a pervasive and persistent phenomenon in Europe during the last two decades. At a general level, the discussion above suggests that this phenomenon reflects similar political forces, namely the political preferences of the majority, consisting of the insiders in the labor market. But there is also a very interesting variation across countries, with regard to the extent of the unemployment problem and the timing and type of policy reforms adopted. Some countries, notably Spain, that introduced tight labor-market restrictions at an early stage, experienced very high unemployment and have only lately introduced reforms in the direction of a two-tier system[21]. In the UK, labor markets were instead deregulated in more conventional ways in the 1980s, by various reforms diminishing the influence of unions. Countries like Sweden introduced legislation providing higher job security in the early 1970s, but avoided high unemployment – for some time, at least – by expanding public employment. Understanding such differences in policy reform is an important topic for future research.

Another interesting question is why different countries resort to different combinations of firing protection and unemployment insurance to protect the insiders against the risk of becoming unemployed. Buti, Pench and Sestito (1998) point out that in cross-country data, there is a negative relationship between these two policies: countries, such as Italy, where firing is very difficult also tend to have very small unemployment-insurance programs, and vice versa. In the previous subsection, we discussed some comparative-static results, relating equilibrium unemployment insurance to exogenous hiring and firing rates. But what makes countries choose different combinations of these instruments? One possible answer is related to the political influence of the insiders: firing protections are of more benefit to the currently employed, while unemployment insurance is of more benefit to the currently unemployed. Thus, the combination of these two tools that is chosen probably reflects the relative political influence for the insiders. But to more thoroughly address this issue, we must go beyond the simple median-voter model discussed so far, and investigate other sources of political influence. Labor unions in many countries are very well organized and

[21] Recent US development towards two-tier labor contracts have been characterized by differences in wages, rather than in job security. This difference may relate to the oft-noted difference in wage flexibility on the two sides of the Atlantic (we owe this observation to Alan Auerbach).

well-connected with political parties on the left. Moreover, their political activities go well beyond the voting behavior of their members. Such activities take us into the domain of special-interest politics, however, which is the topic of Part II below.

Finally, another first-order question is to understand why European and US (more generally Anglo-Saxon) labor markets differ to such an extent. An interesting possibility is that we observe a manifestation of multiple equilibria. The simple model in this section includes a two-way mapping: from labor-market policy to unemployment and from unemployment to policy. Suppose it was enriched with, say, a search model of the labor market, so that equilibrium unemployment was explicitly determined by maximizing choices of firms and workers. It is not inconceivable that such a model would allow for multiple equilibria with different levels of unemployment being supported by different equilibrium labor-market policies, in analogy with the aforementioned paper by Hassler et al. (1998).

4.4. Notes on the literature

A huge literature discusses how exogenous economic policy affects unemployment – see the recent surveys by Bertola (1998), Layard and Nickell (1999) and Mortensen and Pissarides (1999). Research on what mechanisms determine the economic policies that have impact on the labor market is, however, much more scant. The model of voting over unemployment insurance of Subsections 4.1 and 4.2 draws on Wright (1986). It can be extended to allow self-insurance through borrowing and lending, as in Hassler and Rodriguez Mora (1999), or to allow feedback effects from unemployment insurance to equilibrium unemployment, as in Hassler et al. (1998). The political conflict between insiders and outsiders and the issues discussed in Subsection 4.3 have been studied by Saint-Paul (1993, 1996), who also discusses the political feasibility of alternative reforms. Some evidence documenting the conflict between insiders and outsiders over unemployment insurance emerges from the opinion polls of European citizens reported by Boeri, Börsch-Supan and Tabellini (2001).

5. Capital vs. labor

This section has two goals. One is to address a positive question: how is the tax burden split among different tax bases, in particular between labor and capital. According to the basic principles of optimal taxation, labor should be taxed much more highly than capital, as capital is a more elastic tax base. Indeed, in a multi-period context many proponents argue that the optimal steady-state tax rate on capital income is zero[22]. Yet, the observed effective tax rates on capital are positive and often large.

[22] This result is reported in Judd (1985) and Chamley (1986); see also Lucas (1990) and Auerbach and Hines (chapter 21, this volume). It is based on the idea that a capital tax creates a distortion between current and future consumption, which grows with the date of future consumption. With unionized labor markets, however, a labor tax can be as distorting, or even more distorting, than a capital tax; see for instance Daveri and Tabellini (1997).

In a sample of 14 OECD countries, the average effective tax rates on capital and labor over the period 1991–1995 were about the same (about 38%). These measured tax rates vary considerably across countries and over time. In a number of countries, effective tax rates on capital are higher than on labor, even in countries with fairly competitive labor markets, such as the UK and the USA[23]. One simple reason for high taxes on capital is that a majority of the voters prefer them. This result immediately falls out, once we generalize the simple model from Section 2 to include capital as well as labor. Capital income is more concentrated than labor income. Hence, a majority of the voters gain from shifting a larger share of the tax burden to capital, despite the efficiency losses. Another often discussed reason for high taxes on capital is the celebrated capital-levy problem [Fischer (1980)]. The elasticity of already accumulated capital is zero. Hence, sequential policy decisions run into a typical credibility problem and, in equilibrium, capital is taxed even more highly than what is ex-ante optimal for a majority of the voters. Both results are discussed in Subsection 5.2, within the familiar median-voter model with Downsian candidates.

Our second goal in this section is methodological: to explore an alternative model of representative democracy, where candidates are not motivated by the desire of winning the elections per se, but by the desire to implement their own preferred policy. Hence, in Subsection 5.3 we abandon the traditional Downsian model of electoral competition. Instead, we study a representative democracy, where the voters elect outcome-motivated politicians who choose policy once in office. Different candidates represent different ideologies. This setup directs the attention to a new question: who is chosen by the voters to make policy decisions? Voters realize that different political candidates will make different policy choices once in office. A general result is that this way of modelling representative democracy generates *strategic delegation.* The reason is timing: policy choice takes place after the elections, and possibly much later. At the time of the elections, voters realize that policy will be chosen in an environment where the policymaker will face a different set of incentive constraints. To cope with these forthcoming incentive constraints, they find it optimal to elect someone with preferences different from their own. In this setting, following Persson and Tabellini (1994b), strategic delegation allows voters to circumvent the capital-levy problem: the elected policymaker has stronger ex-post incentives to protect accumulated capital than the majority itself. This is just an example, however, and many other instances of strategic delegation have been studied in the literature. We will also return to this theme in Part II. Finally, we extend the model further, and ask whether the suggested equilibrium with an elected *citizen candidate* is in fact consistent with optimal entry

[23] The source is Daveri and Tabellini (1997), who, in turn, extend a methodology formulated by Mendoza, Razin and Tesar (1994) which exploits information on tax income and aggregate tax bases. Effective tax rates on capital from detailed studies of the tax code, using the methodology originally developed by Jorgenson, such as King and Fullerton (1984), often give a very different picture than the "macro" methodology of Mendoza et al.

into the political process. Thereby, we discuss the model of representative democracy proposed by Osborne and Slivinsky (1996) and Besley and Coate (1997).

5.1. A simple model of capital and labor taxation

To deal with capital-formation and credibility problems in a simple way, we extend our simple model from Section 2 to include two time periods. The preferences of the ith individual are:

$$w^i = U(c_1^i) + c_2^i + V(x^i),$$

where the notation follows the previous sections. The labor–leisure choice is only made in period 2. In that period, the individual is thus constrained by the same time constraint (2.1) as before. The period-1 and period-2 budget constraints are:

$$c_1^i + k^i = 1 - e^i, \qquad c_2^i = (1 - \tau_L) l^i + (1 - \tau_K) k^i,$$

where τ_L and τ_K are the tax rates on labor and capital. Both exogenous gross-factor returns have been normalized to unity.

To avoid two-dimensional individual differences, we make the simplifying, but counterfactual, assumption that type i's endowments of initial wealth $1 - e^i$ and of effective time $1 + e^i$ are perfectly negatively correlated. The idiosyncratic parameter e^i thus captures the relative importance of labor and capital in an individual's income. Solving the utility-maximization problem, for given tax rates, we get the labor- and capital-supply functions, which – by the quasi-linear preferences – only depend on the "own tax rate":

$$l^i = L(\tau_L) + e^i \tag{5.1}$$

$$k^i = K(\tau_K) - e^i. \tag{5.2}$$

As before, we assume that e^i is distributed with a c.d.f. $F(\cdot)$. For simplicity, we now set the mean to zero: $e = 0$. Since asset income is more concentrated in the population than labor income, it is natural to assume that the median value of e^i, defined by $F(e^m) = \frac{1}{2}$, is positive.

The final piece to complete the model is the government budget constraint:

$$\tau_L L(\tau_L) + \tau_K K(\tau_K) = G. \tag{5.3}$$

For simplicity, we abstract from the use of government revenue in this section and only treat the (per capita) revenue requirement G as a given parameter. We also assume that $\text{Max}[\tau_L L(\tau_L)] > G > 1$: the labor tax base is large enough to finance the whole of G, but the capital tax can never be sufficient for this purpose. This assumption, which could be somewhat relaxed, rules out multiple equilibria in Subsection 5.2.

5.2. Electoral competition between Downsian candidates

In this subsection, we study equilibrium tax policy under the traditional assumption used throughout the first part. Two office-motivated candidates run against each other in a plurality election. Each candidate makes a binding commitment to an electoral platform, namely a vector of tax rates $\boldsymbol{\tau} = (\tau_L, \tau_K)$. In equilibrium, both candidates announce the same policy platform, namely that preferred by the median voter at the time of elections. The voters' preferences hinge crucially on the timing of elections.

5.2.1. Ex-ante elections

We start by assuming that elections take place *at the beginning* of period 1, before private agents have chosen the amount to save in period 1. The platform of the winning candidate is enacted without further re-optimization. A different timing assumption is discussed below.

To characterize the voters' policy preferences, we follow the same approach as in Section 2. Let $W^i(\boldsymbol{\tau})$ be the indirect utility function of individual i:

$$\begin{aligned} W^i(\boldsymbol{\tau}) &= U(1-K(\tau_K)) + V(1-L(\tau_L)) \\ &\quad + (1-\tau_L)L(\tau_L) + (1-\tau_K)K(\tau_K) + (\tau_K - \tau_L)e^i \\ &= W(\boldsymbol{\tau}) + (\tau_K - \tau_L)e^i. \end{aligned}$$

Then, maximize this function with regard to the two tax rates, subject to the government budget constraint and the supply functions defined above. Combining the resulting first-order conditions, we get:

$$\frac{K(\tau_K^i) - e^i}{K(\tau_K^i)}\left[1 + \frac{\tau_L^i}{1-\tau_L^i}\eta_L(\tau_L^i)\right] = \frac{L(\tau_L^i) + e^i}{L(\tau_L^i)}\left[1 + \frac{\tau_K^i}{1-\tau_K^i}\eta_K(\tau_K^i)\right], \tag{5.4}$$

where $\eta_y(x) \equiv \frac{\mathrm{d}y}{\mathrm{d}x}\frac{x}{y} < 0$ denotes the elasticity of y with regard to x. Together with the government budget constraint (5.3), this condition defines the tax policy $\boldsymbol{\tau}^i$ preferred by voter i.

The individual thus wants taxes to be set according to a modified "Ramsey Rule". Consider first the policy preferred by the individual with average relative income from labor and capital. This policy also has some normative appeal; due to quasi-linear preferences, it coincides with the utilitarian optimum. Clearly, with $e^i = e = 0$, the condition reduces to the familiar inverse elasticity formula of optimal commodity taxation, showing that capital should indeed be taxed more lightly than labor, if its supply is more elastic. Intuitively, the average individual does not care about redistribution, only about efficiency: thus his favored tax policy just minimizes the deadweight loss associated with taxation. We refer to this Ramsey policy as $\boldsymbol{\tau}^*$.

When $e^i \neq 0$, redistributive preferences modify this pure efficiency condition in a predictable way. That is, individuals with more labor than capital income ($e^i > 0$)

want the tax rate on capital to be higher and the rate on labor income to be lower, and vice versa if $e^i < 0$ (recall that elasticities are defined to be negative):

$$\begin{matrix} \tau_K^i \lessgtr \tau_K^* \\ \tau_L^i \gtrless \tau_L^* \end{matrix} \quad \text{as} \quad e^i \gtrless 0. \tag{5.5}$$

The monotonicity of these preferences implies that $\boldsymbol{\tau}^m$, the tax policy preferred by the median voter with endowment e^m, is a unique Condorcet winner[24]. As $e^m > 0$, the implied equilibrium tax policy $\boldsymbol{\tau}^m$ has a higher taxation of capital and a lower taxation of labor than our normative benchmark policy $\boldsymbol{\tau}^*$. In this sense, there is thus overtaxation of capital, due to the skewed distribution of wealth, which implies that the pivotal voter relies relatively more on labor income than on capital income.

5.2.2. Ex-post elections

Next, suppose that elections are held at *the end* of period 1, after the savings decision has been made. This is the case of tax policy under "discretion", discussed by Fischer (1980) and Persson and Tabellini (1990). Under this assumption, agents still behave according to Equations (5.1)–(5.2) in their economic decisions, except that the expected tax rate on capital replaces the actual tax rate in the savings function. Their voting behavior is now different, however, as it takes place once aggregate capital is given. Hence, we refer to their policy preferences as *ex post.*

To describe these preferences, note that when elections are held, the elasticity of capital with regard to the *actual* tax rate is zero: $\eta_K(\tau_K) = 0$. The capital stock depends on the expected tax rate; once the capital is in place, changing τ_K does not further reduce it. With this in mind, consider the average voter with $e^i = 0$. This voter has no stake in redistribution and only cares about efficiency. He would like to tax capital as highly as possible (the inelastic factor), so as to reduce the distorting tax on labor (the elastic factor). Thus, his ex-post optimal policy is $\tau_K = 1$, for any aggregate capital stock inherited from the past[25]. For a "laborer", with $e^i > 0$, the redistributive motive reinforces these incentives for ex-post expropriation. Thus, since $e^m > 0$, a majority of the voters wants to set $\tau_K = 1$ for any outstanding capital stock.

It follows that this tax policy is announced by both candidates in their electoral platforms. As this is perfectly foreseen when the savings decision is made, nobody

[24] The monotonicity follows because we can write agents' indirect utility in this model as

$$W^i(\boldsymbol{\tau}) = W(\boldsymbol{\tau}) + (\tau_L - \tau_K)e^i,$$

which is linear in the idiosyncratic parameter e^i. (The linearity property is not destroyed if we substitute the government budget constraint into this expression.)

[25] Recall our previous assumption that the capital tax base cannot be large enough to finance the whole of G.

saves anything. In equilibrium, $k = 0$ and all the revenue must be raised by taxing labor alone: $\tau_L L(\tau_L) = G$ [26]. This is the classical "capital-levy problem": low taxes on capital are not credible to investors foreseeing the outcome in the subsequent political equilibrium. In our simple model, this problem manifests itself in a disastrous equilibrium: that is, a tax policy with a confiscatory capital tax, which gives individuals no incentive to save.

Clearly, this prediction is too strong. Even though we observe higher taxes on capital than those prescribed by simple optimal taxation models, we rarely observe confiscatory rates. The literature has suggested a number of reasons why credibility problems might not have such drastic consequences. These include reputational effects, linking future expected taxation to current taxation, and the possibility for agents to protect their capital ex post, by tax avoidance or capital flight. All these forces mitigate the credibility problem in capital taxation, but do not necessarily entirely remove the problem. Thus, lack of credibility may compound the overtaxation of capital for purely political reasons.

5.3. *Equilibrium taxation with citizen candidates*

Up to this point, we have retained two crucial assumptions about the political process. First, candidates are office-motivated: they only care about winning the election per se. Second, they can make binding promises ahead of the elections. Neither assumption is very palatable. It is hard to justify the assumption of binding electoral promises: policy decisions are made once in office, without being constrained by promises made during the electoral campaign. Moreover, politicians often have their own political agenda, their ideology or view of the world, which motivates their policy decisions once in office. In this subsection, we consider a different model of the political process, based on alternative assumptions. Politicians are directly motivated by policy outcomes; they are "citizen candidates". That is, each candidate for political office is just an ordinary individual in society who – like everybody else – is solely motivated by her utility function. Moreover, tax policy is chosen after the election, once in office. This means that pre-election announcements by political candidates are never credible. Voters are forward looking, and select among candidates on the basis of their "ideology", correctly predicting that an elected candidate will simply set the ex post optimal policy.

We first follow Persson and Tabellini (1994b) by showing that this kind of environment naturally invites the voters to resolve credibility problems in capital taxation via strategic delegation. We then discuss another important aspect of the political process, to which this approach naturally directs attention: the entry stage on the political arena.

[26] For lower values of G such that $G < \text{Max}_{\tau_K}[\tau_K K(\tau_K)]$, there are better equilibria in a social-welfare sense. But, unlike the prescription of the Ramsey Rule, these also have very unbalanced taxes, with all the revenue being raised by the capital tax and the labor tax set at $\tau_L = 0$. These other equilibria are discussed by Persson and Tabellini (1999a).

Here, we borrow from Besley and Coate (1997) and show that such strategic delegation is indeed an equilibrium – though not the only one – in a game with endogenous and costly entry by citizen candidates.

5.3.1. Preferences over candidates

Assume that the prospective policymaker is one of the individuals in the model, uniquely identified by her endowment e^P, were P stands for policymaker[27]. The timing of elections is also crucial in this setting. We assume that elections are held at the start of period 1, before the savings decision[28]. But policy is set at the end of period 1, after the elections and after capital has been accumulated.

At this point in time, any elected policymaker maximizes her ex-post utility, taking into account that, since capital is already in place, $\eta_K(\tau_K) = 0$. Thus, as discussed in the previous subsection, any elected policymaker with $e^P \geqslant 0$ finds it optimal to set $\tau_K = 1$ for all k. A policymaker with $e^P < 0$, however, behaves differently. He still perceives $\eta_K(\tau_K) = 0$, and this pushes him to set a high τ_K. But the redistributive motive pulls him in the opposite direction. He is at an interior optimum and his preferences for τ_K can be obtained from the modified Ramsey rule in Equation (5.4), by setting $\eta_K(\tau_K) = 0$. We denote this interior optimum capital tax rate, as a function of e^P and k, by $T(e^P, k)$. It is implicitly defined by Equation (5.4) with $\eta_K(\tau_K) = 0$ and by the government budget constraint (5.3). It can be shown that this function has partials $T_e, T_k > 0$. Intuitively, the higher is the average tax base k, the greater is the efficiency gain from taxing it; conversely, the lower is e^P algebraically, the greater is capital income relative to labor income for policymaker P and hence, the lower is his desired capital tax rate.

We can summarize the above discussion as follows. The tax rates enacted by policymaker P, if elected and given outstanding capital k, are defined by

$$\begin{aligned} \tau_K^P &= \begin{cases} 1 & \text{for} \quad e^P \geqslant 0, \\ T(e^P, k) & \text{for} \quad e^P < 0, \end{cases} \\ \tau_L^P L(\tau_L^P) &= \begin{cases} G - k & \text{for} \quad e^P \geq 0, \\ G - T(e^P, k)k & \text{for} \quad e^P < 0. \end{cases} \end{aligned} \tag{5.6}$$

Clearly, these preferences imply a monotonic relation between an elected office-holder's endowment e^P and his chosen tax policy. Call this mapping $\boldsymbol{\tau}(e^P)$. This mapping is known and understood by voters and investors at the time of elections. Seeing who wins the elections, investors correctly anticipate the forthcoming tax policy and invest accordingly. Voters also take this into account when they vote: they realize that electing a policymaker with a high value of e^P discourages investment through

[27] What we call the policymaker endowment can also be interpreted as reflecting her ideology on a left to right scale.

[28] If elections were held ex post, after the capital accumulation decision, nothing essential would change relative to the Downsian equilibrium. This case is thus ignored.

the expectation of high capital tax rates. And conversely, a policymaker with a low value of e^P is a credible signal that τ_K will be kept low.

More precisely, the voters' ex ante preferences over tax rates map one for one into preferences over policymakers. Specifically, the utility of voter i is given by $w^i = W^i(\boldsymbol{\tau}(e^P))$. And the policymaker preferred by individual i is

$$e^{iP} = \arg\max_{e^P} \left[W^i(\boldsymbol{\tau}(e^P))\right].$$

Given the assumed election timing, e^{iP} is the policymaker type who finds it ex post optimal to carry out the ex ante optimal policy of individual i. Such policy, denoted $\boldsymbol{\tau}^i = (\tau_K^i, \tau_L^i)$, is implicitly defined by Equations (5.4) and (5.3). Thus, knowing τ_K^i, we can implicitly find e^{iP} from the expressions

$$\tau_K^i = T(e^{iP}, K(\tau_K^i)), \qquad \tau_L^i L(\tau_L^i) = G - \tau_K^i K(\tau_K^i). \tag{5.7}$$

Recall that $T(\cdot)$ is strictly increasing in e^P only for $e^P < 0$, while the ex ante optimal tax rate on capital for voter i, i.e. τ_K^i, is increasing over the whole range of e^i. Several conclusions follow. First, every voter prefers a policymaker who relies more on capital than on herself – i.e., $e^{iP} < e^i$. Indeed, every voter prefers a policymaker in the minority of the population, with $e^i \leqslant 0$. "Right-wing" candidates thus have a natural advantage in this setting, as they more credibly protect capital from overtaxation out of self-interest. Second, the induced preferences over policymaker types are themselves monotonic in voter type.

In other words, when the electoral horizon is long enough, there is a motive for strategic delegation: to protect capital from expropriation, the majority elects a policymaker with higher capital income than average. Indeed, by the monotonicity established above, the policymaker, e^{mP}, preferred by the median voter, e^m, is the unique Condorcet winner in the population; i.e., this is the only candidate who would win a pair-wise contest against any other candidate. It is thus tempting to argue, as in Persson and Tabellini (1994b), that the election of e^{mP} and the ex post implementation of τ^m, that is, the median voter's ex ante optimal policy, is the equilibrium outcome. This argument is not complete, though. Why would e^{mP} find it optimal to run as a candidate? This candidate would also rather have somebody else set policy, given the credibility problem. To know whether e^{mP} running and getting elected is an equilibrium, we really must study an explicit prior stage, where political candidates enter the competition.

5.3.2. *Endogenous entry of candidates*

Let us thus assume that the ex ante elections-cum-policy game is preceded by an entry decision of prospective political candidates. With this addition, the game has the following stages. (i) Any individual (of any type e^i) in the population can decide to run as a candidate, at a cost (in terms of second-period consumption) of ε.

(ii) An election is held among those running as candidates; whoever receives a plurality of the vote wins, any tie is resolved by tossing a coin. (iii) Individuals make their savings decisions. (iv) The elected candidate chooses a tax policy $\boldsymbol{\tau}$; if no candidate had decided to run, a default policy $\overline{\boldsymbol{\tau}}$ is implemented. (v) Individuals make their labor-supply decision.

From the discussion above, we already know how to characterize the outcome from stages (iii)–(v). At stage (ii), each individual anticipates this outcome and votes for the candidate maximizing her expected utility, given the vote of other individuals[29]. At stage (i) – again anticipating the outcome at the following stages – an individual chooses to enter only if this gives her higher expected utility than not entering, given the entry decision of other individuals.

We now adapt the results in Besley and Coate (1997) to this model and characterize its different equilibria.

5.3.2.1. Single-candidate equilibria. We have already argued that the policymaker type e^{mP} preferred by the median voter e^m is the unique Condorcet winner among potential candidates. Thus e^{mP} is assured to win against any other candidate if he decides to run. But if he runs, no alternative candidate $e^{P'}$ will ever find it worthwhile to incur the entry cost of running; this would not in any way affect the policy outcome and hence not the utility of $e^{P'}$, who would thus only bear the cost of running. This means that whenever e^{mP} runs in equilibrium, he must run as an uncontested candidate. The condition for such an equilibrium to exist is

$$W^{mP}(\boldsymbol{\tau}(e^{mP})) - W^{mP}(\overline{\boldsymbol{\tau}}) \geqslant \varepsilon. \tag{5.8}$$

The condition says that the utility gain, relative to the default policy, for e^{mP} from running and choosing her ex-post optimal policy must outweigh the cost of running. It is obviously fulfilled if the default policy $\overline{\boldsymbol{\tau}}$ is sufficiently far from the equilibrium policy $\boldsymbol{\tau}(e^{mP})$ or else if the running cost is small. As there is no gain from holding office per se, no second candidate of type e^{mP} has an incentive to run, as he would just incur the entry cost without influencing policy[30]. In this equilibrium, the majority thus succeeds in completely resolving the credibility problem by strategic delegation to a "right-wing" policymaker, who is assured to win the election and has the right ex-post incentives to implement the majority's preferred ex ante policy $\boldsymbol{\tau}^m$.

5.3.2.2. Two-candidate equilibria. Even though candidates occasionally run uncontested in majoritarian electoral systems, single-candidate races are not very common. We therefore study the conditions for equilibria with two candidates: e^R and e^L, say.

[29] We also rule out weakly dominated voting strategies. Together, these conditions imply sincere voting in one- and two-candidate elections. For a more careful discussion, see Besley and Coate (1997).

[30] This suggests a free-rider problem among the e^{mP} types. One can also add exogenous benefits from holding office, as do Osborne and Slivinsky (1996).

Intuitively, this requires that e^R finds it worthwhile to run, given that e^L is running, and vice versa. As in the single-candidate case, this calculation involves trading off the influence on policy against the entry cost. But it also requires that each candidate stands some chance of winning. In our setting with monotonic candidate preferences, this means that an individual with median policy preferences e^m is indifferent between the two candidates. In this event, the two candidates have the same chance of winning. Formally, sufficient conditions for a two-candidate equilibrium are

$$W^m(\boldsymbol{\tau}(e^R)) = W^m(\boldsymbol{\tau}(e^L)),$$
$$\tfrac{1}{2}\left[W^I(\boldsymbol{\tau}(e^I)) - W^I(\boldsymbol{\tau}(e^J))\right] > \varepsilon, \quad I, J = R, L, \quad I \neq J.$$

In this kind of equilibrium, two candidates with endowments on opposite sides of the median voter's preferred type $e^{m\,P}$ are running against each other. Each of them has an incentive to enter so as to balance the other candidate, provided that their endowments are different enough (otherwise, a fifty–fifty chance of winning does not offset the cost of running). It follows that there are many different two-candidate equilibria. In each of these, a right-wing candidate with $e^R < e^{m\,P}$ balances a left-wing candidate with $e^L > e^{m\,P}$ at the same utility distance from the median voter's preferred policy. All voters with endowments $e^i < e^m$ vote for e^R, whereas all voters with $e^i > e^m$ vote for e^L. These equilibrium voting strategies keep a third intermediate candidate from entering. But, as Besley and Coate (1997) discuss, there may also be equilibria with three or more candidates entering.

5.3.3. *Discussion*

In the two-candidate equilibria studied above, the voters only succeed in delegating the credibility problem in an "expected sense", for once elected, the two candidates will pursue different policies. This feature illustrates a general property of the citizen-candidate model; equilibrium policy in two-candidate elections does not converge to the policy preferred by the median voter, given that such a voter exists. This contrasts starkly with the prediction of the Downsian model with office-motivated candidates. It also contrasts with models of electoral competition, where outcome-motivated candidates who commit to policy platforms ahead of the elections converge to the policy preferred by the median voter – cf. Wittman (1977, 1983). But as Alesina (1988) argues, once we assume candidates to be outcome-motivated, the common assumption of commitment becomes very strong, as it clashes with the elected candidate's ex-post incentives. If the commitment assumption is relaxed, policy convergence requires some reputational mechanism fostering long-run cooperation between the candidates.

In general, the citizen-candidate model provides a general-equilibrium approach to politico-economic modelling. It is attractive in the sense that it starts from primitives. Assumptions are only made about the individuals' preferences, endowments and technologies, and about the institutions of economic and political interaction. This makes it possible to use the model for a clean welfare analysis of political

equilibria. Furthermore, the citizen-candidate model can handle situations where a simple median-voter approach would fail: citizen-candidate equilibria exist under very general conditions, including many conditions where no Condorcet winner exists. We have thus not made the model full justice, by applying it in such a smooth setting.

The citizen-candidate approach is not without its weaknesses, however. The main lesson from this approach is precisely the importance of the entry stage in the political arena. The lack of pre-existing electoral candidates, however, makes it hard to introduce political parties in the analysis[31]. Moreover, multiple equilibria make it hard to use the model for generating testable hypotheses. Nevertheless, the citizen-candidate model is an ingenious construction that offers an interesting alternative for modelling of electoral equilibria.

Finally, the model focuses on representative democracy, but, to date, its applications have neglected the agency problems arising when the candidates' policy preferences are loosely defined, or when endogenous rents from office can motivate policy decisions. In Part III, we return to such agency problems. Their investigation in a setting with citizen-candidates is a difficult, but interesting, topic for future research.

5.4. Notes on the literature

There is a large literature on the so-called "capital-levy" problem, starting with Fischer (1980). This literature investigates the implications of lack of credibility in capital taxation, as well as how to restore credibility through reputation or institutional design. Persson and Tabellini (1990, 1999a) provide extensive surveys of credibility problems in fiscal and monetary policy, while Persson and Tabellini (1995) survey the literature on capital taxation and tax competition. The theories of optimal taxation are surveyed by Auerbach and Hines (2002, Chapter 21 of this volume).

The modeling of politicians as ideological – or outcome-motivated – individuals at least dates back to the work of Hibbs (1977) and Wittman (1977); see also Wittman (1983). Alesina (1988) relaxed the commitment assumption in a setting with rational voters, and showed that policy convergence no longer applies.

The idea that strategic delegation allows the principal to cope with incentive constraints on the agent was first applied in industrial organization by Vickers (1985) and Fershtman and Judd (1987). It found several other natural applications in political economics, with regard to credibility problems in monetary policy [Rogoff (1985)] and capital taxation [Persson and Tabellini (1994b)), international policy coordination [Persson and Tabellini (1992)], hierarchical decision making by different levels of government inside a federation (Persson and Tabellini (1996)], bargaining inside a legislature [Chari, Jones and Marimon (1997), Besley and Coate (1998a) – see also Subsection 10.2 in Part II below].

[31] Recent work by Rivière (1998) and Besley and Coate (1998b) does attempt to introduce parties in a citizen-candidate setting.

The citizen-candidate model was formulated by Osborne and Slivinsky (1996) and Besley and Coate (1997). It has since been applied to analyze several economic policy problems; references are given in the sections to follow. Besley and Coate (1998b) includes a penetrating discussion of the efficiency properties of political equilibria in citizen-candidate models.

Part II. Special-interest politics

Many economic policy decisions create concentrated benefits for a few well-defined groups, with the cost diffused in society at large. This occurs in public finance, but also in trade policy and regulation. Whenever economic policy benefits narrowly define special interests, the political incentives to influence the design of such policies are much stronger for the beneficiaries than for the majority bearing the cost. A classical example of this systematic bias is agriculture. Farmers in all democracies are generously supported through trade policies, direct subsidies, and various other support programs. Several explanations have been suggested for this phenomenon. Many stress that farmers have more homogenous economic interests and therefore find it easier to get organized. Others emphasize that farmers are less ideologically biased and therefore become a natural target for politicians who vie for electoral support. Some also point out that farmers are concentrated in rural electoral districts which are often overrepresented in legislatures, or that legislators representing rural interests often hold important positions as ministers or chairmen of congressional committees.

The public-choice literature has emphasized one of these mechanisms in particular. Due to their higher stakes in the various programs, beneficiaries are more likely to get politically organized, whereas the interests of the unorganized general public are neglected. This idea dates back to the work of Schattschneider (1935), Tullock (1959), Olson (1965), Weingast, Shepsle and Johnsen (1981), Becker (1983, 1985) and several others. Mueller (1989, 1997) includes excellent surveys of the earlier literature. More recent contributions have focused on *structural* models of the political process, trying to identify specific features of the political system that confer power to some groups rather than others, or that entail systematic biases in aggregate spending. In this part, we survey some of these recent contributions. The main results are theoretical. Compared to Part I, we focus more on analytical methodology and less on specific empirical applications.

Multi-dimensionality of policy space renders the simple electoral approach adopted in Part I useless, as it would unavoidably result in Condorcet cycles. To predict likely policy outcomes – and, in particular, which groups are most powerful in the struggle for benefits – we must specify the institutional details of the policy process. Different branches of political economics have taken this route in recent years, specifying the policymaking process as an extensive-form game and assuming rational individual behavior. Some of the empirical implications are not very different from those of earlier public-choice literature. The older approach often lacked micro-political

foundations, however, relying instead on non-derived influence functions, political support functions, or vote functions. Contributors to the more recent literature have tried to fill this gap, by being more explicit on the institutional assumptions and more uncompromising on the requirements of individual rationality.

To illustrate the effects of the different political determinants of policy, we stick to the same economic example throughout[32]. We choose a very simple model, which highlights the more general phenomenon of concentrated benefits and dispersed costs in a transparent way. Thus, we study a society where the government uses a common pool of tax revenues to provide an array of publicly provided goods, the benefits of which are completely concentrated to well-defined groups. The most important question motivating the analysis concerns the allocation among groups: which groups are politically powerful and how is this related to political institutions?

In Section 6, we formulate the basic model and derive some benchmark allocations. In the subsequent sections, we apply three different state-of-the-art models to our policy example. Each one of these studies a specific feature of the political process in detail.

In Section 7, we formulate a *legislative bargaining* model, developed by researchers in American congressional politics, to study decision-making rules and budgetary procedures. Political power reflects the assignment of agenda-setting or amendment rights, and the sequencing of decisions. Institutions that centralize decision-making power by conferring strong proposal rights and limiting amendments induce a small size of government, but distort the allocation in favor of whoever holds such powers.

In Section 8, we use a model of *lobbying* as common agency, developed by researchers in trade policy, to study the influence activities of organized interest groups. The model directs the attention to campaign contributions and the organizational pattern of interest groups. Groups organized as a lobby have disproportionate influence on the final allocation, which generally results in suboptimal allocations. If taxpayers are less politically organized than the beneficiaries of the spending programs, because they have smaller stakes individually, a large government emerges.

In Section 9, we use a model of *electoral competition*, developed by public economists, to study the electoral platforms chosen by two vote-maximizing candidates. This is the model of probabilistic voting which was introduced already in the Introduction to Part I. As in Part I, the candidates are office-motivated and commit to policy platforms ahead of the elections. When choosing which party to support, however, voters trade off these economic policy platforms against predetermined ideological preferences. Political power reflects the distribution of voters' ideological preferences across groups; more powerful groups include a large number of "swing-voters", i.e., voters who are mobile across candidates because they do not care too much about ideology. To win the elections, both candidates direct economic benefits towards these a-ideological voters.

[32] The treatment in this part extends a survey along similar lines in Persson (1998).

While these approaches yield useful insights, each of them still gives a partial answer to the question which are the most powerful groups. A formal integration of the different approaches is only beginning to take shape. Section 10 discusses the main results so far. We start by studying the interaction between *elections and lobbying*: office-seeking politicians use the lobbying revenues to influence voters. Next, we illustrate the interaction between *legislative bargaining and elections*: voters in each of multiple voting districts elect outcome-motivated politicians as their representatives in a subsequent legislative-bargaining game. Finally, we study the interaction between *legislation and lobbying*: different lobbies seek to influence finance-motivated politicians involved in legislative bargaining to confer benefits on their groups. The results do not always constitute a convex combination of the results from partial models[33].

Overall, the results in this part remove us very far from the median-voter outcome of Part I. Politics is much more than just vote counting. To understand the political determinants of policy, we must pay attention to many fine details of the political process. But the research we survey is mainly theoretical. It needs to be better integrated with empirical work, to gain a more complete understanding of the relative importance of each of these details.

6. A simple model

Consider a society with $\mathcal{J}$ distinct groups of identical individuals. Group $J = 1, \ldots, \mathcal{J}$ has size (mass) N^J, $\sum_J N^J = N$, where N is the size of the entire population. Individuals in group J have the quasi-linear preferences

$$w^J = c^J + H(g^J), \tag{6.1}$$

where c^J denotes the consumption of private goods (the same for every group member) and g^J is the per-capita supply of a publicly provided good. The increasing and concave function $H(\cdot)$, with $H(0) = 0$, is thus defined over a (private) good, which benefits group J only and must be publicly provided in an equal amount per capita (we could easily add some externalities onto other groups, at the cost of additional algebraic complexity). Individual income is the same in all groups: $y^J = y$. A unit of income (private consumption) can be costlessly converted into one unit of any of the $\mathcal{J}$ publicly provided goods, and taxation is lump-sum. This model can be interpreted

[33] An important omission is that we entirely disregard bueraucratic behavior and its interaction with other parts of the political process. Economists have recently built structural models of the interaction between interest groups and the bureacracy to study regulatory capture [Laffont and Tirole (1993)], and political scientists have studied the legislature's control of bureacracy [McCubbins, Noll and Weingast (1987)].

in a number of ways: groups could be defined by their preferences, occupation, age or other personal attributes, or by geographical location.

6.1. A normative benchmark

As a normative benchmark, consider the utilitarian optimum, obtained by maximizing the Benthamite welfare function, $\sum_J \frac{N^J}{N} w^J$, subject to the resource constraint $\sum_J N^J(g^J + c^J) = Ny$. The resulting benchmark allocation is pretty obvious, namely to set the vector $\boldsymbol{g} \equiv (g^J)$ such that the average marginal benefit in each group equals the marginal social cost of unity:

$$H_g(g^*) = 1. \tag{6.2}$$

For future reference, we denote aggregate spending associated with this allocation by $G^* = Ng^*$.

This allocation could easily be implemented if each of the group-specific goods were financed by group-specific lump-sum taxes, (τ^J), so that: $c^J = y - \tau^J = y - g^J$. If full decentralization of spending and financing to each group were feasible, this would be the optimal institutional arrangement. The policymakers' incentives would not be distorted, and the socially optimal policy would emerge as an equilibrium.

In the real world, however, it is often impossible to design the tax system so that the taxpayers' financing of a group-specific good precisely coincides with the beneficiaries. For instance, the beneficiaries may be identified by their personal attributes or occupation, and not by residence; or, else, their individual characteristics may be unobservable, as in the case of preferences.

Our goal in this part is to explore the incentive problems arising under centralized financing, and how different political institutions change these incentives and the resulting allocations. Thus, throughout, we retain the stark but simplifying assumption that all publicly provided goods must be financed out of a common pool of tax revenues, with equal contributions from each group. The policy instruments are always the same: the vector $\boldsymbol{g} \equiv (g^J)$ of publicly provided group-specific goods and a common lump-sum tax, τ, and they are always subject to the same government budget constraint: $N\tau = \sum_J N^J g^J \equiv G$, where G, as above, denotes aggregate expenditures.

In this set-up, individuals have distorted incentives and there is sharp disagreement over policy. The reason is that the cost of financing the public good is shared between the groups. Hence, beneficiaries would like to over-spend on their preferred public good, since the cost of providing this good is shared with others. Conversely, every group wishes to reduce spending on the other public goods, since they do not internalize any benefit from them.

Adding externalities, so that the local public good g^J also affects the utility of groups different from J adds other considerations, but does not remove the incentive problems discussed throughout this part. Even if full decentralization was feasible, it would not deliver the social optimum, as the externalities would not be

internalized. Under full centralization, the incentive problems due to cost sharing would remain, as long as different groups preferred different combinations of public goods. For simplicity, throughout this part, we thus neglect externalities.

6.2. *The basic common-pool problem*

To illustrate these incentive problems, we start with a simple decision-making procedure. Each group decides freely on the supply of the public good, whereas the tax rate is determined residually. Individual utility in group J can then be written as

$$W^J(\boldsymbol{g}) = y - \tau + H(g^J) = y - \sum_I g^I \frac{N^I}{N} + H(g^J). \tag{6.3}$$

An equilibrium is a vector $\boldsymbol{g}^D$ (where the superscript D stands for decentralized spending), such that each group J maximizes $W^J(\boldsymbol{g})$ with respect to g^J, taking equilibrium expenditures by all other groups as given. It is straightforward to verify that equilibrium spending here satisfies

$$H_g(g^{J,D}) = \frac{N^J}{N}. \tag{6.4}$$

Since the right-hand side of Equation (6.4) is less than 1, all groups overspend compared to the social optimum: $g^{J,D} > g^*$ for all J, and smaller groups overspend to a larger extent. This is the familiar "common-pool" problem: each group fully internalizes the benefit of its own public good, but – as financing is shared – it internalizes only the fraction N^J/N of the social marginal cost of higher taxes. The problem here lies in the collective choice procedure, where the tax rate is residually determined once all spending decisions have been made in a decentralized fashion. Concentration of benefits and dispersion of costs lead to excessive spending when residually financed out of a common pool of tax revenue.

Even though the nature of the problem is evident, the remedy of full decentralization of financing may be difficult to enforce. As mentioned above, it may be hard to adapt the system of financing to the relevant group structure. Common-pool problems thus arise in many situations. For instance, they can be due to lack of information, so that some spending decisions must be decentralized to local governments, government agencies, or public enterprises, while financing remains centralized. Moreover, the incentive problem illustrated above does not disappear under fully centralized decisions on spending, as each group will seek to influence the central government to satisfy its own interests. Concentration of benefits and dispersion of costs imply that each group retains an incentive to demand an oversupply of goods to its own group, and an undersupply to the other groups, to avoid paying high taxes. Which groups will be most politically powerful, in taking advantage of this opportunity, depends on group attributes but also on political and budgetary institutions. The

remaining sections discuss how the policy problem is resolved in alternative settings and how these settings shape observed policy outcomes.

6.3. Notes on the literature

This section draws on Persson and Tabellini (1994c). Models of this sort have been used extensively to discuss incentive problems in local public finance and to contrast alternative budgetary procedures. In particular, Besley and Coate (1998a), Lockwood (1998) and Daveri (1998) consider a similar set-up, but assume that local public goods have externalities on other groups. They contrast decentralized and centralized arrangements, pointing to a trade-off between two opposite incentive problems. Centralization makes it more likely that spillover effects are internalized, but cost sharing generates the incentive problems discussed throughout this part. Full decentralization on the other hand, prevents the externalities from being internalized. The preferred institutional arrangement thus depends on which of these incentive problems is the worst.

When there is a vertical hierarchy of decision makers, as with federal and local governments, lack of commitment by the principal may induce a "soft budget constraint" on the agent. As common-pool problems, soft-budget-constraint problems may lead to overspending. Dewatripont and Maskin (1995) is the classical reference on soft budget constraints in a principal–agent set-up. Qian and Roland (1998) and Bordignon, Manasse and Tabellini (2001) have studied versions of this problem in local public finance.

7. Legislative bargaining

A large empirical literature has studied how budgetary institutions correlate with fiscal outcomes. Most of this literature focuses on intertemporal fiscal policy choices, however. Cross-sectional comparisons suggest that specific procedures are associated with smaller budget deficits. In particular, centralization of budgetary power to the prime minister or the finance minister, two-stage budgeting with prior setting of deficit targets, restrictions on amendments of spending proposals, and constitutional limits on deficit spending, seem to promote more fiscal discipline[34]. Less attention has been devoted to implications of alternative budgetary procedures for the size of government, with a few exceptions noted below. This is an unfortunate omission, as one of the underlying problems which "stricter" budgetary procedures are supposed to solve, namely the common-pool problem, may also distort the level of spending.

As noted in the previous section, the problem stems from excessive decentralization of spending: each group is the arbiter of spending on its own local public good. In this

[34] In the USA, a procedure similar to giving power to the Treasury is to require all spending proposals to be channelled through one committee; see Cogan (1994).

section, we analyze a centralized procedure: the policy vector $(\boldsymbol{g}, t)$ is now assumed to entail spending on geographical districts. To be implemented, a policy must be approved by a majority of districts, according to specific procedural rules. If there is no agreement, a default outcome – the status quo – kicks in. The model of this section thus purports to describe decision making in a legislature, and the rules capture stylized features of the budget process. We draw on the seminal work by Romer and Rosenthal (1978) on agenda setting. Their analysis was, in turn, extended by Baron and Ferejohn (1989), whose legislative-bargaining framework has become a work-horse model for the analyses of the American Congress and other legislatures. We ask how bargaining power is determined inside the legislature, and how alternative procedures shape aggregate spending[35].

7.1. A simple legislative-bargaining model

Groups are distinguished by their geographical location and each location is represented by one member in the legislature. This representative is "outcome motivated" and is a perfect delegate of her constituency, in that her preferences are of the same form as in Equation (6.3). The number of districts and representatives $\mathcal{J}$ is now assumed to be odd, with $\mathcal{J} \geqslant 3$. These assumptions fit well the system of representation in the American congress with plurality elections in multiple single-member districts. Interpretations more fitting to parliamentary systems with proportional representation are also possible, but less straightforward.

The "budget process" in a legislative session consists of the following sequence of events: (i) One of the representatives, $J = a$, is chosen to be the agenda setter[36]. (ii) Representative a makes a policy proposal, $\boldsymbol{g}$. (iii) The legislature votes on the proposal. If a simple majority approves the proposal – that is, at least $\frac{\mathcal{J}-1}{2}$ other legislators vote in favor – then $\boldsymbol{g}$ is implemented (a always votes for her own proposal). If not, a status-quo outcome, $\overline{\boldsymbol{g}} = (\overline{g}^J) : \ \overline{\tau} = \sum \frac{N^J}{N} \overline{g}^J$, is implemented.

In the jargon of the legislative-bargaining literature, we are thus considering a *closed rule* – i.e., proposals cannot be amended – with only one round of proposals. Amendments and multiple rounds, with proposal rights alternating between legislators, are discussed below.

7.2. Political equilibrium

Consider first the choices by legislators $J \neq a$ at the voting stage (iii). Clearly, any legislator will only approve proposals $\boldsymbol{g}$ which, from her own point of view, are not

[35] Baron (1993) has applied the legislative-bargaining model to a similar policy problem.

[36] We do not model the criteria for selecting the agenda setter. In real-world democracies this choice presumably reflects electoral outcomes. But very few papers have tried to model this formally; see, however, McKelvey and Riezman (1991) and the discussion at the end of Subsection 10.2.

worse than the status quo (we assume that indifferent legislators always vote yes to a proposal). From Equation (6.3) and the definition of $\bar{\boldsymbol{g}}$, legislator $J \neq a$ votes in favor of $\boldsymbol{g}$ if

$$W^J(\boldsymbol{g}) - W^J(\bar{\boldsymbol{g}}) = H(g^J) - H(\bar{g}^J) - \sum_I \frac{N^I}{N}(g^I - \bar{g}^I) \geqslant 0. \tag{7.1}$$

Consider next the proposal stage (ii). Here, the agenda setter maximizes her own pay-off, given by Equation (6.3), subject to the government budget constraint, the "incentive compatibility constraints" (7.1) holding for a majority coalition $\mathcal{M}$, including at least $\frac{\mathcal{J}-1}{2}$ other legislators, and the non-negativity constraints $g^J \geqslant 0$ for all J. Eliminating the multipliers from the Kuhn–Tucker conditions to this problem and manipulating the solution, we can write the following conditions describing the equilibrium proposal, denoted with a superscript B:

$$\begin{aligned}
&H_g(g^{J,B}) = \frac{N^J}{N}\frac{1}{1 - \sum_{I \in \mathcal{M}} \frac{N^I}{N}\frac{1}{H_g(g^{I,B})}}, \qquad J = a,\\
&g^{J,B} = 0, \quad J \notin \mathcal{M},\\
&H(g^{J,B}) - H(\bar{g}^J) = \sum_{I \in \mathcal{M}} \frac{N^I}{N}(g^{I,B} - \bar{g}^J), \qquad J \neq a,\ J \in \mathcal{M},\\
&|\mathcal{M}| = \frac{\mathcal{J}-1}{2}.
\end{aligned} \tag{7.2}$$

To understand this equilibrium, consider the incentives of a. To get support from other legislators, a must spend costly tax revenue in their districts. We can consider a's problem in two stages. In the first stage, she minimizes the tax rate τ necessary for obtaining support for every value of g^a, implying an increasing function $T(g^a)$. The cost-minimization stage basically involves minimizing the term $\sum_{I \in \mathcal{M}} \frac{N^I}{N}\frac{1}{H_g(g^I)}$ in the denominator of the first right-hand-side expression in Equation (7.2). Given this "cost function", she then simply maximizes $H(g^a) + y - T(g^a)$ in the second stage, with respect to g^a. This has several consequences:

(1) A version of Riker's (1962) so-called size principle will hold: a chooses a *minimum winning coalition*, $\mathcal{M}$, which is composed of $\frac{\mathcal{J}-1}{2}$ other legislators. All districts outside the winning coalition get no spending at all, even though they bear the cost of taxes.

(2) For the members of $\mathcal{M}$, a spends only as much as necessary to get their vote (i.e., to satisfy condition (7.1) with equality), leaving them as well off as with the default policy.

(3) The minimum winning coalition is composed of those legislators whose support is cheapest to obtain. These are the legislators with the lowest default pay-offs, $\bar{g}^J$. A weak status-quo position may thus be to the advantage of a legislator and her

district. Even though a district with a weak position gets less public goods, when its legislator is part of $\mathcal{M}$, the chance of being part of the majority is higher, the weaker is that position. In a richer model where legislators also differ in the relative weight attached to private versus public consumption, the majority would include the legislators who care *more* about public consumption, since their vote is cheaper to buy – this point is once more discussed in Section 10. Finally, ceteris paribus, size – or rather misrepresentation in the districting – matters. As their legislator still has one vote, it is cheaper for a to please districts with a smaller number of voters. They are thus more likely to be included in the majority.

(4) The resulting allocation is asymmetric and suboptimal compared to the utilitarian benchmark. Districts not in $\mathcal{M}$, certainly get less (namely zero) spending than in the utilitarian optimum. Whether the members of the majority get more or less, depends on parameters and on the shape of $H(\cdot)$. As long as the default allocations $\overline{g}^J$ of the majority districts are not too high, however, they will typically get less: $g^{J,B} < g^*$ for $J \neq a$, $J \in \mathcal{M}$. Under these circumstances, district a certainly gets more: $g^{a,B} > g^*$. To show this formally, rewrite the first row of Equation (7.2) as

$$H_g(g^{a,B}) - 1 = -\frac{\lambda_{\mathcal{N}} + \sum_{I \in \mathcal{M}} \frac{N^I}{N}\left(1 - \frac{1}{H_g(g^{I,B})}\right)}{1 - \sum_{I \in \mathcal{M}} \frac{N^I}{N} \frac{1}{H_g(g^{I,B})}},$$

where the left-hand side is the expression defining the utilitarian optimum. Thus, the right-hand side measures the deviation from the efficiency benchmark. Note that the first term in the numerator, $\lambda_{\mathcal{N}} \equiv \sum_{I \notin \mathcal{M}} \frac{N^I}{N}$, is the population share of the districts not belonging to the majority. As the second term in the numerator is also positive, given $H_g(g^{J,B}) - 1 > 0$ for $J \in \mathcal{M}$, overprovision to district a follows. Furthermore, the overprovision to a is larger, the smaller is the population share of the majority (i.e., the larger is $\lambda_{\mathcal{N}}$), as this reduces the cost of expanding g^a while compensating the legislators in the majority. The asymmetry also depends on the default positions; the lower is the average value of $\overline{g}^J$, the more powerful is the agenda setter. Since $\overline{g}^J$ refers to the status quo if the new legislation is voted down, this suggests that we should observe more asymmetric benefits for certain types of government programs. Specifically, infrastructure projects – where the natural status quo is no projects – should be more asymmetrically distributed across groups than entitlement programs – where the natural status quo is the existing policy (and where beneficiaries are probably also more evenly distributed across voting districts).

(5) Finally, whether the model predicts aggregate overspending or not depends on parameters and on the concavity of $H(\cdot)$, and there is no presumption that the

bias goes either way[37]. But this model contains two useful lessons for the design of budgetary procedures. First, aggregate spending is more likely to be low, the smaller are the default outcomes in $\bar{g}$. If the status quo entails little spending, as with zero-base budgeting, the strong agenda-setting powers of one legislator discipline all the others. Second, suppose that different legislators differ in their valuation of public vs. private spending, and that agenda-setting power is given to a legislator who spends little for his constituency or – thinking of bargaining within government – to a minister without portfolio, such as the finance or Treasury minister. Then, the agenda setter does not expand his preferred public good, and concentration of proposal power delivers small aggregate spending.

The political-science literature has discussed other reasons for conferring strong agenda-setting powers on some legislators, besides control of aggregate spending. All legislatures necessarily display some division of labor across issues, due to the need to split the work load, as well as the varying background of legislators. Giving control over certain issues to some individuals provides incentives to invest in issue-specific competence and information gathering. In the US congress, for instance, this specialization and control is manifested in powerful standing committees with considerable agenda-setting powers over the issues under their jurisdiction[38]. Standing committees are also found in parliamentary systems, although in such systems the ministries have many of the corresponding agenda-setting tasks. Other political scientists have instead taken a more sanguine view, emphasizing that a particular organization of Congress facilitates the legislators' desire to earn re-election by conferring the benefits of pork-barrel programs to their districts[39]. The model thus captures something important: real-world legislatures are organized in a way that makes some representatives more powerful than others over certain issues, a power which influences the allocation of spending.

7.3. Extensions

Power associated with proposal rights is, however, modified and diluted in several ways by the procedures adopted in real-world legislatures. One mechanism is the amendment right of other legislators, another is separation of proposal powers: different legislators have agenda-setting rights over different policy dimensions. We briefly discuss these in turn.

[37] The flatter is H_g, the more likely is over-spending. Consider the special case when $\mathcal{J} = 3$, such that the majority $\mathcal{M}$ consists of a single legislator m. Furthermore assume that $\bar{g}^J = 0$ and $H(g^J) = \alpha[\ln(g^J)]$. We then get $g^a = 3\alpha - e$, $g^m = e$, and thus $G = 3\alpha = G^*$. Thus the allocation of spending is distorted, with $g^a > g^m$ if $\alpha > \frac{2e}{3}$, and $g^m > g^*$ if $\alpha < e$ (where e is the base of the natural logarithm). But the aggregate level of spending coincides with the utilitarian optimum.

[38] An informational view on legislative organization, including the rationale for vesting agenda-setting powers with legislators and committees, has been emphasized by some political scientists; this is well exposed in Krehbiel (1991).

[39] See, for instance, Weingast and Marshall (1988).

7.3.1. Amendment rights

Instead of the closed rule analyzed earlier, assume now an open rule, according to which the initial proposal can be amended by some other legislator. It is common practice to pitch an offered amendment against the initial proposal in a vote, and then to either allow a new round of amendments to the winning proposal, or else pitch the winning proposal against a default policy. Including such amendment rights in the model above diminishes the gains that a could expect from equilibrium policy. As the amendment right allows the amender to tilt the proposal in her own favor, albeit at the cost of legislative delay, any initial proposal must make a majority of the legislators better off, not only relative to the default outcome but also relative to their continuation value from further bargaining. Baron and Ferejohn (1989) and Baron (1993) demonstrate that equilibrium policy generally entails more equally distributed benefits under open rule than under closed rule. Although the precise results depend on the details of the amendment procedure, equilibria may, in some cases, come close to implementing the efficient solution. These models have an infinite horizon, however, and to simplify, the size of government is exogenously given. As far as we know, no theoretical result exists on how amendment rights shape aggregate spending.

A related model is due to Lockwood (1998), who adapts previous results by McKelvey (1986) and Ferejohn, Fiorina and McKelvey (1987) to a setting similar to ours. The legislature must choose how many projects of a given size to activate. Different projects benefit different legislators, and can have externalities on other districts. Financing is shared among all districts. Legislative rules are as follows. First, each legislator makes a proposal. These proposals are then randomly ordered into an agenda, and are voted on sequentially. Finally, the winning proposal is voted against the status quo. This procedure insures that an equilibrium exists and is unique, even if there is no Condorcet winner. If externalities are weak or negative, only a bare majority of the projects are funded; these are the projects with the lowest cost. If externalities are strongly positive, on the other hand, a larger number of projects is funded. Moreover, which projects are funded reflects the costs and the externalities, but not the intensity of preferences of individual legislators with regard to their favorite projects. Thus, this procedure does not guarantee an egalitarian outcome, but it reduces the importance of particularistic political preferences.

7.3.2. Separation of budgetary powers

Many existing legislatures split the budgetary procedures into two stages: first, aggregate spending is determined, to be followed by allocative decisions. It is often argued that this two-stage budgeting insulates the decision on aggregate spending from the special-interest politics that disrupts incentives, and that this leads to better aggregate decisions[40]. We now investigate whether this is true in our simple model.

[40] See, for instance, Von Hagen (1998).

For simplicity, assume that $\mathcal{J} = 3$, and that all groups are of equal size: $N^J = \frac{N}{3}$. Suppose that the budgetary procedure involves two stages. In the first stage the legislature decides on overall spending G, or – equivalently – on the common tax rate $\tau = G/3$. This decision is taken by a single majority under a closed rule, after a proposal by agenda-setting legislator a_τ. A defeated proposal results in default aggregate spending, $\overline{G}$. In the second stage, a different agenda-setter, $a_g \neq a_\tau$, makes an allocation proposal, subject to $\sum_J g^J = G$, with G given from the first stage. If this proposal is defeated, the first-stage budget is split according to a simple sharing rule $\overline{g}^J = \frac{1}{3}G$, where the assumption of equal sharing is made for simplicity. The status quo for aggregate spending in the second stage is the *equilibrium* outcome from the first stage.

The second-stage equilibrium is simple. To get the necessary majority, agenda-setter a_g must propose to spend enough in one of the other districts, say m_g, to just exceed the status-quo outcome: $g^{m_g} = \frac{1}{3}G$. She spends nothing in the minority district, n_g, and allocates the remaining budget to her own district: $g^{a_g} = \frac{2}{3}G$. As the total budget and the tax rate are already fixed, taxes do not enter the allocation decision. The allocation distortion remains, but we are now mostly interested in the level of spending.

The first-stage outcome depends on who makes the proposal and whether the composition of the second-stage majority is known. Suppose first that the first-stage proposal is made by a member of the future majority, and that her identity is known. Thus we have: $a_\tau = m_g \neq a_g$. The optimal level of G for the first-stage proposer maximizes $[H(g^{m_g}) - \frac{1}{3}G] = [H(\frac{1}{3}G) - \frac{1}{3}G]$, and satisfies

$$G^{m_g} = 3H_g^{-1}(1).$$

Thus G^{m_g} coincides with our benchmark optimum G^*. The intuition is simple: at the first stage, m_g internalizes the full benefits to her own district of aggregate spending, and these are equal to a third of the social benefits. As she also internalizes a third of the social costs (her district's share of the tax bill), she faces the right marginal incentives when it comes to aggregate spending[41]. If the future majority composition is indeed known, G^* always collects a majority against $\overline{G}$. Interestingly, if $G^* > \overline{G}$, a_g supports this because she wants as high a revenue as possible to allocate at the second stage. A stable majority thus suggests the two parts of the budget. If instead the status quo involves aggregate "overspending" $G^* < \overline{G}$, a_τ instead gets support of n_g, the minority legislator at the next stage, who has an obvious incentive to keep aggregate spending and taxes down.

In parliamentary systems, there is indeed a presumption that majorities are predictable; a point discussed in more detail in Part III. But without further institutional

[41] Naturally, the allocative distortion remains, and thus nothing insures that G^* is still optimal, given that allocative distortion.

detail, nothing pins down the second-stage majority. Therefore, consider an alternative case, where $a_\tau \neq a_g$, but a_τ is only part of the future majority with 50% probability. In this case, the optimal level of G, from the point of view of a_τ, maximizes $[\frac{1}{2}H(\frac{1}{3}G) - \frac{1}{3}G]$, namely

$$G^n = 3H_g^{-1}(2).$$

Clearly, $G^n < G^{m_g} = G^*$. When the first-stage proposer is not certain of being a "residual claimant" on the second-stage budget, she has a stronger interest in keeping down the size of the budget. A similar point is encountered in Part III. The desirability of such separation of powers in the political system is perhaps not obvious in the present setting. But separation of powers can unambiguously play to the voters' advantage, once we introduce agency problems.

We conclude this section with a general remark. Most of the work in the legislative-bargaining literature is quite partial in that it takes the preferences of the legislature as given. Where do the outcome-oriented preferences of legislators come from? Legislators' behavior may also be influenced by other motives, such as a desire to raise funds, to get re-elected, or to use political power for their own private agenda creating an agency problem vis-a-vis the voters. If lobbies and voters understand these motives and how the legislative process works, would they not adapt their behavior to influence the policy outcome? To answer questions of this kind, we must obviously leave partial models behind and study interactions between different aspects in the political process. Section 10 gives different examples of such interactions, while Part III deals with agency problems. But first, we turn to the partial models of lobbying and voting.

7.4. Notes on the literature

The formal literature on extensive-form games of collective choice dates back to the pioneering work of Shepsle (1979) on structure-induced equilibria and Romer and Rosenthal (1979) on agenda-setting powers. Models of legislative bargaining were first formulated by Baron and Ferejohn (1989) in an infinite-horizon cake splitting problem, and applied to the provision of local public goods by Baron (1991, 1993). A different extensive-form game, allowing for amendments in a particular way, was studied by McKelvey (1986) and Ferejohn, Fiorina and McKelvey (1987); its applications to public finance are yet to be explored [cf., however, Lockwood (1998)].

Sequential budgeting has been studied in different settings. Von Hagen (1998) discusses it in a more comprehensive analysis of budgetary procedures. Persson, Roland and Tabellini (1997) discuss the benefits of two-stage budgeting coupled with strong agenda-setting powers in a model of agency. Their point is dealt with again in Part III. Ferejohn and Krehbiel (1987) analyze a median-voter model with sequential voting in different dimensions, and argue that two-stage budgeting may fail to deliver the alleged benefits; but their set-up does not entail a common-pool problem.

A large empirical literature compares alternative budgetary institutions across political systems. It has dealt with European countries [von Hagen (1992), von

Hagen and Harden (1994)], Latin America [Alesina et al. (1999), Inter-American Development Bank (1997)], and the US states [Alesina and Bayoumi (1996), Poterba (1994), Bohn and Inman (1996)]. This literature indicates that specific procedures are associated with smaller budget deficits. The correlation with the size or composition of spending has not been much discussed, except by Kontopoulos and Perotti (1997, 1999). Poterba and von Hagen (1999) contains a number of contributions on budgetary procedures and fiscal performance.

8. Lobbying

Our next model of policymaking focuses on the influence or lobbying activities of interest groups. Policy decisions are here assumed to be centralized in the hands of a semi-benevolent government. But the government can be influenced by organized interest groups. How does this influence activity modify the allocation and level of government spending? Which groups are likely to be favored? Recent rational-choice oriented analyses have focused either on the incentives for lobbies to gather information and provide it to the policymakers, or else on their influence-seeking activities. In the latter tradition, Grossman and Helpman (1994, 1995) and several others have adapted the common-agency model of Bernheim and Whinston (1986) to something of a work-horse model of lobbying, which has been used for studying trade policy, commodity taxation and other policies. Here, we follow Persson (1998) in applying the common-agency model to the study of group-specific government spending[42].

8.1. A simple lobbying model

As Olson (1965) noted a long time ago, influence activities entail a free-rider problem: all members of a group benefit, irrespective of whether or not they contribute to the lobbying. Some groups are successful in overcoming this free-rider problem, others are not. We follow the literature by not modelling how this takes place, and just assume that a subset $\mathcal{L}$ of groups are organized to influence public-goods allocation in their favor. Thus, we study a policy game with two stages. (i) Each lobby J non-cooperatively and simultaneously presents their common agent, "the government", with a per capita contribution schedule $C^J(\boldsymbol{g})$, giving a binding promise of payment, conditional on the chosen policy. The objective of the lobby is to maximize the *net* welfare of its members, namely $N^J(W^J(\boldsymbol{g}) - C^J(\boldsymbol{g}))$, where $W^J(\boldsymbol{g})$ denotes the welfare from the

[42] Persson and Tabellini (1994c) study local public-goods provision in a common-agency model, but impose unappealing restrictions on the strategies used by interest groups.

economic policies, as defined in Equation (6.3). (ii) The government sets $\boldsymbol{g}$ so as to maximize a weighted sum of social welfare and contributions:

$$W(\boldsymbol{g}) = \eta \sum_J N^J W^J(\boldsymbol{g}) + (1-\eta) \sum_{J \in \mathcal{L}} N^J C^J(\boldsymbol{g}), \tag{8.1}$$

where η $[0 \leqslant \eta \leqslant 1]$, is a measure of the government's benevolence.

An equilibrium of the game is a Subgame perfect Nash equilibrium in the contribution schedules and the chosen policy vector. Following the literature, we shall confine ourselves to equilibria in (globally) truthful contribution schedules, namely those satisfying

$$C^J(\boldsymbol{g}) = \text{Max}[W^J(\boldsymbol{g}) - b^J, 0], \tag{8.2}$$

where b^J is a constant set optimally by the lobby[43].

8.2. *Political equilibrium*

To derive an equilibrium in truthful strategies, we can exploit its property of being jointly Pareto optimal for the government and each and every lobby. The equilibrium vector $\boldsymbol{g}$ will therefore maximize the sum of the net welfare of the organized lobbies, $\sum_{J \in \mathcal{L}} N^J(W^J(\boldsymbol{g}) - C^J(\boldsymbol{g}))$, and the government objective $W(\boldsymbol{g})$, component by component. Using the definitions above, it is thus as if the optimal policy maximizes the weighted sum

$$\eta \sum_{J \notin \mathcal{L}} N^J W^J(\boldsymbol{g}) + \sum_{J \in \mathcal{L}} N^J W^J(\boldsymbol{g}), \tag{8.3}$$

where aggregate welfare for the non-organized groups is defined in the same way as in Equation (6.3). In other words, the equilibrium coincides with the solution to a planning problem, where the non-organized groups are underweighted relative to the organized groups, to an extent that depends on the government's benevolence. The first-order conditions to Equation (8.3), defining the equilibrium allocation, denoted with a superscript L, can be rewritten as

$$\begin{aligned} &H_g(g^{J,L}) - 1 = -(1-\lambda_{\mathcal{L}})(1-\eta) \leqslant 0, \quad J \in \mathcal{L}, \\ &H_g(g^{J,L}) - 1 = \lambda_{\mathcal{L}}(1-\eta)/\eta \geqslant 0, \quad J \notin \mathcal{L}, \end{aligned} \tag{8.4}$$

where $\lambda_{\mathcal{L}} = \sum_{J \in \mathcal{L}} \frac{N^J}{N}$ is the share of the population organized in a lobby. The left-hand side of Equation (8.4) is the expression defining the utilitarian optimum, so the

[43] A (locally) truthful contribution schedule has the property that $\partial C^J(\boldsymbol{g})/\partial g^I = \partial W^J(\boldsymbol{g})/\partial g^I$ for any I and everywhere. That is, the slope of the contribution schedule in any direction is equal to the true marginal benefit of the policy in that direction for lobby J. See Grossman and Helpman (1994) and Dixit, Grossman and Helpman (1997) for further details and for a discussion of the restriction to truthful strategies in common-agency games.

right-hand side measures the deviation from the optimum benchmark. Several results are apparent:

(1) As is evident from Equation (8.4), the equilibrium can be socially optimal: $\boldsymbol{g}^L = \boldsymbol{g}^*$. Unsurprisingly, this happens when $\eta = 1$, so that the government is completely benevolent and does not value contributions at all, or when $\lambda_{\mathcal{L}} = 0$, with no contributing groups to worry about. But it also happens when $\lambda_{\mathcal{L}} = 1$, when everyone belongs to a lobby. Stated otherwise, suboptimal policies are only enacted due to incomplete participation in lobbying. The reason is that each group has a strong incentive to lobby, not only for large g^J, for itself, but also for low provision to other groups, to pay lower taxes. When all groups are organized, they offset each other's influence. Since they reveal their marginal preferences to the government by their truthful contributions, the true marginal social cost is correctly internalized in the policy decision.

(2) Generally, however, public consumption is misallocated: organized groups get more and unorganized groups less than the optimal amount. Intuitively, overprovision to the organized lobbying groups is larger if the government values contributions more (η is smaller) and hence pays more attention to the preferences expressed by the lobbies. If $\eta \to 0$, the government only cares about contributions, and provision to the unorganized groups also goes to zero. The overprovision is also larger, the lower is the share of the organized groups (the lower is $\lambda_{\mathcal{L}}$), as the lobbies – and indirectly the government – then internalize a smaller share of the social marginal costs. Note, however, that only the combined size of the organized lobbies influences the outcome; large and small organized groups obtain as much support per capita. Clearly, our implicit assumption that all members of each group belong to the lobby is driving this result.

(3) There is no presumption of aggregate overprovision. While there is certainly overprovision to the organized groups, there is underprovision to the non-organized ones. Not only do the preferences of the non-organized receive a smaller weight in the policy decision, but the tax burden of provision to non-organized groups is internalized by organized groups, which communicate this to the government. In a richer model, with individual heterogeneity over the preferences for private versus public consumption, it is plausible that lobbies would consist of individuals with a high preference for the publicly provided goods. The reason is that they have a higher stake on the policy outcome, and, hence, are more likely to overcome the free-rider problem of getting organized. The intuition why consumers are underrepresented in lobbying is familiar from games over trade policy. In this event, it is easy to show that lobbying results in aggregate overspending compared to the normative benchmark.

Finally, this model can be adapted to also include the choice over a global public good, which benefits all groups in the same way. In this case, it easily shown that the provision of this public good is not distorted by lobbying. Intuitively, lobbying induces the government to underweigh the welfare of unorganized individuals. But these individuals are affected by the national public good just like anyone else, both

as taxpayers and as beneficiaries. With enough symmetry, neglecting their welfare does not distort the policy choice. The general lesson is that lobbying distorts policy, which has a different impact on different groups, as in our case of local public goods.

The common-agency model of lobbying aggregates the influence activities of many interest groups into a policy decision, in an elegant and simple way. It also sheds light on how the pattern of organization across groups shapes the policy outcome. But the model leaves some crucial issues aside. On the one hand, one lacks a precise model of the process whereby groups get politically organized and others not. This is a difficult question, to which there is still no satisfactory answer. The asymmetries driving the misallocation of public goods must thus be assumed, or defended on empirical grounds, rather than explained. On the other hand, the "government" and the process of policy choice is still a black box. If the lobbying model captures what goes on between elections, what exactly does the objective function in Equation (8.1) capture? It is really impossible to answer this question without a structural model of policy choice. Following Grossman and Helpman (1996), we embed a lobbying into the electoral framework of the next section in Subsection 10.1, and show that the parameter η can then be derived from more structural assumptions. In Subsection 10.3, we also combine lobbying and legislative bargaining.

8.3. Notes on the literature

Austen-Smith (1997) gives a recent survey of the literature on lobbying, while Mueller (1989) surveys the older literature. An influential branch of the literature, not discussed here, approaches lobbying as strategic transmission of asymmetrically held information; see Potters and van Winden (1992) and Austen-Smith and Wright (1992). Grossman and Helpman (1994) were the first to use Bernheim and Whinston's (1986) common-agency approach to model lobbying in the case of trade policy. Dixit (1996b) applies the same approach to commodity taxation, showing why the well-known Diamond–Mirrlees production-efficiency prescription would almost surely be violated in political equilibrium. Aidt (1998) adopts it in analyzing environmental taxes. Dixit, Grossman and Helpman (1997) contains a general discussion of the common-agency approach with applications to public finance. Boylan (1995) points to the similarities between this approach and the literature on auctions.

Grossman and Helpman (2001) give an extensive overview of the recent literature on interest groups and their influence on economic policymaking.

9. Electoral competition

We have seen how the ability of interest groups to get organized into lobbies and be represented by powerful legislators gives them an edge in the struggle for policy benefits. But some groups may also have particular attributes, in their role as voters, which make them an attractive target for office-motivated politicians. Our

last partial model of centralized policymaking and special-interest politics therefore focuses on electoral competition. There is no lobbying, no legislative bargaining, and no separation of decisions on spending and taxes. Policy decisions are made by two competing candidates who maximize the probability of winning the election. They make binding promises of policy favors to interest groups ahead of the elections. Unlike in Part I, the two candidates are not identical, and different voters have "ideological preferences" for one or the other. At the time of elections, these ideological preferences are traded off against the announced economic policy benefits. When announcing policy favors, the candidates take into account which groups are more likely to be swayed. The question we ask is which groups have most influence on electoral promises.

The modeling in this section follows Lindbeck and Weibull (1987) and, subsequently, Dixit and Londregan (1996) who modified the probabilistic voting model of Enelow and Hinich (1982) and others from a spatial setting to redistribution among groups. In this section, we adapt their models – which both deal with direct income redistribution out of a given budget – to our policy problem with group-specific public consumption out of an endogenous pool of tax revenue.

9.1. A simple model of electoral competition

Consider the model of Section 7, but add two office-motivated political parties $P = L, R$. Before the election, both parties non-cooperatively commit themselves to specific policy platforms, $\boldsymbol{g}_L$ and $\boldsymbol{g}_R$. Parties also differ in another dimension, unrelated to the announced economic policies – we shall refer to this dimension as "ideology", although it could also involve other features, such as the personal characteristics of the party's leadership. This ideological dimension is a permanent attribute of each party, in the sense that it cannot be changed at will during the electoral campaign.

This ideological difference among parties is reflected in the voters' preferences: each voter has an "ideological bias" for or against party L. Specifically, member i of group J has the extended utility function

$$v^{iJ} = \kappa^J W^J(\boldsymbol{g}) + (\sigma^{iJ} + \theta)D^L, \tag{9.1}$$

where D^L takes a value of unity if party L wins the election and zero otherwise. Further, σ^{iJ} is an *individual-specific* parameter, κ^J is a *group-specific* parameter, and θ is a random variable capturing the party preferences of the *whole population*. Thus, individuals are distinguished by two features: the group they belong to, indexed by J, and their individual party bias, σ^{iJ}. Individuals with $\sigma^{iJ} > 0$ (< 0) have a bias in favor of (against) party L, which is stronger the greater is σ^{iJ} (in absolute value). Individual party bias is distributed within each group according to a uniform distribution on the interval $[-\frac{1}{2s^J}, \frac{1}{2s^J}]$. That is, the distribution of σ^{iJ} for all i belonging to group J has density s^J. Thus, each group has members inherently biased towards each of the parties, even though the distribution of party bias differs across groups. Moreover,

groups also differ in the strength of their ideological motives; the larger is the parameter κ^J, the more all the individuals in J care about economic well-being relative to ideology. Finally, the random variable θ captures the average popularity of party L in the population as a whole. We assume that θ has a uniform distribution on $[-\frac{1}{2h}, \frac{1}{2h}]$. The realization of θ is unknown to the parties when announcing their policy platforms, so that the election outcome is uncertain from their point of view.

Equations (6.3) and (9.1) imply that voters in group J supporting party R all have $\sigma^{iJ} < \kappa^J[W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L)] - \theta$. Let us identify the "swing voter" in group J as the voter who – given the parties' platforms – is indifferent between the two parties. We denote these voters' party bias as $\sigma^J(\boldsymbol{g}_R, \boldsymbol{g}_L, \theta)$, defined by:

$$\sigma^J(\boldsymbol{g}_R, \boldsymbol{g}_L, \theta) \equiv \kappa^J \left[W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L)\right] - \theta. \tag{9.2}$$

Swing voters toss a coin when deciding how to vote.

9.2. Political equilibrium

The two parties simultaneously and non-cooperatively choose their platforms, so as to maximize the probability of winning the election[44]. To specify the party objectives, first note that the distributional assumptions allow us to write the vote share of party R as

$$\pi^R = \sum_J \frac{N^J}{N} s^J \left[\sigma^J(\boldsymbol{g}_R, \boldsymbol{g}_L, \theta) + \frac{1}{2s^J}\right].$$

By definition of σ^J in Equation (9.2) and the assumption that θ is uniformly distributed with density h, its probability of winning can be written as

$$p^R = \text{Prob}\left[\pi^R \geqslant \tfrac{1}{2}\right] = \tfrac{1}{2} + h\left[\sum_J \frac{N^J}{N}\frac{s^J}{s}\kappa^J \left[W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L)\right]\right], \tag{9.3}$$

where $s \equiv \sum \frac{N^J}{N} s^J$ is the average density of party bias across groups. Party R sets its platform so as to maximize this expression, subject to the budget constraint. As the probability of winning for party L is given by $1 - p^R$, as $\boldsymbol{g}_L$ affects p^R symmetrically but with the opposite sign as $\boldsymbol{g}_R$, and as the two parties face the same budget constraint, they face the same decision problem. Specifically, this optimization problem does not include any party-specific variables. It should thus come as no surprise that a Nash equilibrium involves identical policy platforms $\boldsymbol{g}_L = \boldsymbol{g}_R$. By Equation (9.2), this

[44] The Nash equilibrium obtained if parties maximize their vote share is identical [see Lindbeck and Weibull (1987) and Dixit and Londregan (1996)]. In this case, the random variable θ could be omitted from the model.

implies $\sigma^J(\boldsymbol{g}_R, \boldsymbol{g}_L, \theta) = -\theta$. As the expected value of θ is zero, each party is doing its best to capture the votes of the ideologically neutral voters in each group, namely those with $\sigma^{iJ} = 0$.

In view of this, the first-order conditions determining the allocation of equilibrium spending across groups can be written as

$$\frac{N^J}{N}\frac{s^J}{s}\kappa^J H_g(g^J) - \frac{N^J}{N}\sum_I \frac{N^I}{N}\frac{s^I}{s}\kappa^I = 0. \tag{9.4}$$

The equilibrium thus entails a generalized Hotelling-type result. Despite the multidimensional policy space, the two parties converge on the same platforms. The intuition for this is simple: the parties compete for the same voters and thus are both trying to buy the electoral support from the same marginal voters in each group. Furthermore, they have the same technology for converting money into expected votes. As a result, it is the distribution of voters' preferences alone that decides the unique equilibrium election outcome.

To characterize equilibrium spending, $\boldsymbol{g}^E$, it is useful to rewrite Equation (9.4) as

$$H_g(g^{J,E}) - 1 = \frac{\sum_I \frac{N^I}{N} s^I \kappa^I - s^J \kappa^J}{s^J \kappa^J}. \tag{9.5}$$

As in the previous two sections, deviations from the utilitarian optimum are determined by the expression on the right-hand side of the equation. A number of insights emerge:

(1) In a politically homogenous society, where the ideological bias is the same across groups – i.e., the densities s^J and the parameters κ^J coincide for all J – electoral competition implements the utilitarian optimum: $\boldsymbol{g}^E = \boldsymbol{g}^*$. This is intuitive: as both parties try to buy expected votes by influencing the voters' marginal utility, their marginal incentives are identical to those emanating from a utilitarian objective, if each group is identical as concerns how easily their vote can be swayed. This result is well known from the literature on probabilistic voting in a spatial setting; it was first demonstrated by Coughlin and Nitzan (1981).

(2) The political clout of a specific group J is conveniently summarized by the term $s^J\kappa^J$. If this term is higher than the weighted average of the other groups, the right-hand side of Equation (9.5) is negative, implying $g^{J,E} > g^*$. The term s^J measures the density of ideologically neutral voters, that is, of voters who only care about economic policies. These are the most mobile voters, and both parties want to please them. The larger is the density of these "swing voters" within group J, the greater is the expenditure directed towards this group. The parameter κ^J instead reflects to what extent voters in group i care about economic well-being as opposed to ideology. Groups who care less about ideology (i.e., groups with a greater κ^J) are favored, since their voters are more mobile. If these features characterize middle-class voters particularly well, the model thus confirms what Stigler (1970)

minted as "Director's Law", namely that redistributive policies will generally favor the middle class. Conversely, groups caring a great deal about ideology and groups with few swing voters lose out, because buying a large number of expected votes in those groups is too expensive[45].

(3) Group size does not play a role in determining political clout. On the one hand, a large group has many voters and is therefore an attractive target for vote buying. On the other hand, it is more expensive to pay for the votes of a large group. As the expression in Equation (9.4) shows, these two effects cancel each other out. Note, however, that we have assumed that parties maximize the probability of winning, taken over the whole population. Thus, we can consider this an implicit assumption of an electoral system with strict proportional representation.

(4) There is no first-order bias in total spending relative to the utilitarian optimum. As Equation (9.5) shows, some groups get more while others get less. The effect on total spending depends in a complicated way on the interplay between political clout, relative group size, and the concavity of the $H(\cdot)$ function. Intuitively, spending is entirely "supply determined" by the two political parties. The presence of a latent common-pool problem with incentives to expand spending at the group level does not influence the outcome, as each party – in its attempt to buy votes from all groups – properly internalizes the aggregate budget constraint.

The analysis can be extended and modified in a number of directions. In the papers by Lindbeck and Weibull (1987) and Dixit and Londregan (1996), direct income transfers support the private consumption of each group. Poorer groups systematically obtain more support, ceteris paribus, as their marginal utility of income is higher (as it would be for a benevolent planner). The same would apply here with a concave utility of private consumption; poor voters would be more hurt by common taxes and need to be compensated with more public consumption. Strömberg (1998) lets groups differ in their turnout rates, denoted as t^J. The political clout of group J in the model above becomes $t^J s^J \kappa^J$. Groups with higher turnout rates would thus get more support. The "transaction costs" in buying votes may also differ systematically across groups. If these costs or the uncertainties in vote buying are lower among the groups belonging to the party's core supporters (because transfers can be more precisely targeted), this may become a counterweight to a strong party bias and rationalize so-called "machine politics", where parties give more favors to their traditional support groups, as discussed in the model by Cox and McCubbins (1986). Dixit and Londregan (1998) study a more general model where parties and voters also have some ideological concerns about income distribution. This allows them to endogenously derive the result that groups composed of middle-class voters are likely to have most electoral clout.

[45] A more general formulation of the model would have the idiosyncratic parameters $\sigma^{i,J}$ distributed according to general group-specific c.d.f. $S^J(\cdot)$, with different means. In this case, the relevant density would be $s^J(0)$, and groups with an ideological bias (a mean far from 0) would lose out, as they would have few ideologically neutral voters.

The model certainly highlights important aspects of how special interests may be favored by parties in their election campaigns. But it also leaves out important aspects of policy making. For one, there is no interest-group activity; each group is just a target for the politicians, and their members just cast their vote like everybody else. For another, the assumption of binding electoral promises is dubious; many policy decisions are made between elections in the running of business by the incumbent government and its administration. Part III discusses how electoral competition might then be played out through retrospective voting.

9.3. Notes on the literature

The probabilistic-voting approach was developed in the spatial-voting model to guarantee existence of equilibrium in situations, such as a multi-dimensional policy space, when a Condorcet winner fails to exist; see Coughlin (1992) for an overview of probabilistic voting and Osborne (1995) for an overview of spatial-voting theory. An adaption of this framework to redistribution among multiple interest groups was made by Lindbeck and Weibull (1987), and their approach was extended by Dixit and Londregan (1996). These papers, and the other papers mentioned in the text, identify a priori the set of interest groups and the group affiliation of each voter. A general treatment of redistribution among ex ante identical voters, resulting from electoral competition between political candidates – without additional attributes – can be found in Myerson (1993b), who derives an equilibrium where each candidate selects a randomized redistribution strategy.

10. Interactions in the political process

So far, we have studied three different models of special-interest politics, each focusing on a separate aspect of political activity. Real-world politics, however, involves a great deal of interaction between these activities. If lobbies or voters understand how decisions are made in the legislature, they will adapt their lobbying behavior or their candidate preferences accordingly. And if electoral platforms systematically favor certain organized groups, they will adapt their campaign contributions accordingly. In the absence of a "grand unified theory"of special-interest politics – a structural model simultaneously encompassing legislation, lobbying and elections – we devote the remainder of this part to the analysis of three simpler, pairwise, forms of interaction.

10.1. Lobbying and elections

The previous model of lobbying is most straightforwardly interpreted as a model of "bribes" to the government. In practice, however, most lobbying takes the form of campaign contributions, either in cash or "in kind", through actions affecting the electoral outcome. We now combine the lobbying model of Section 8 with the voting

model of Section 9, to illustrate how electorally motivated lobbying may influence policy. The central conclusion is that the insights gained in those two sections survive, and carry over to this more general model. Equilibrium policy is influenced by *both* the lobbying activity and the voters' attributes: organized groups, and groups with more swing voters, are over-represented in the political process. Moreover, additional insights are gained about what determines the effectiveness of the lobbies and the size of equilibrium contributions. The analysis is a variant on that in Bennedsen (1998), who in turn extends and simplifies earlier work by Baron (1994) and Grossman and Helpman (1996)[46].

Consider the same model as in Sections 8 and 9, but with some simplifications. Two vote-maximizing parties, L and R, set policy platforms $\boldsymbol{g}_L$ and $\boldsymbol{g}_R$, respectively, in advance of the elections. As before, these parties differ in some "ideological" dimensions. We now assume that all groups are of equal size normalized to unity, such that $\frac{N^J}{N} = \frac{1}{\mathcal{J}}$, and place the same weight on economic outcomes versus ideology, also normalized to unity, $\kappa^J = 1$. Voters in group J still have preferences:

$$v^{iJ} = W^J(\boldsymbol{g}) + (\sigma^{iJ} - \theta)D^L, \tag{10.1}$$

but now θ is given by

$$\theta = \tilde{\theta} + \phi(C_L - C_R).$$

Thus, the average popularity of party L has two components. The term $\tilde{\theta}$ is a random variable, as previously, uniformly distributed on $[-\frac{1}{2h}, \frac{1}{2h}]$. But the overall relative popularity of the two parties is now also influenced by the campaign contributions received by parties L and R, C_L and C_R, respectively. Specifically, voters are biased in favor of the party receiving more contributions, with $\phi > 0$ being a parameter capturing the sensitivity to the difference in campaign spending[47]. This has more than one interpretation: C_L might measure advertising expenditures or media exposure of the leaders of party L, but it might also refer to support actions in favor of L, or against her electoral opponent[48]. As in Section 9, σ^{iJ} is distributed according to group-specific distributions uniform on $\left[-\frac{1}{2s^J}, \frac{1}{2s^J}\right]$ with density s^J.

[46] Riezman and Wilson (1997) study restrictions on contributions in a setting where competing political candidates instead "sell" policies to different interest groups.

[47] Allowing ϕ to differ across groups or indivduals does not matter for the results, since only the average value of ϕ (across groups and individuals) enters the equilibrium expressions. Note that $\phi > 1$ is allowed.

[48] Grossman and Helpman (1996) suggest a slightly different interpretation, which leads to a similar formulation as (10.2). Some voters are fully informed and uninfluenced by campaign contributions. Other voters are uninformed about economic policy platforms, and respond exclusively to campaign contributions. The overall effectiveness of campaign contributions in swaying voters is then related to the frequency of uninformed voters in the population.

By the same logic as previously, the indifferent voter in group J is an individual with preference parameter

$$\sigma^J \equiv W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L) + \phi(C_R - C_L) - \widetilde{\theta}. \tag{10.2}$$

Thus, the identity of this swing voter is affected by campaign spending. All voters in group J with $\sigma^{iJ} > \sigma^J$ prefer party L, all those with $\sigma^{iJ} < \sigma^J$ prefer R. Following the same approach as in Section 9, we can derive the probability of winning for party R as

$$p^R = \tfrac{1}{2} + h\left[\frac{1}{\mathcal{J}}\left[\sum_J \frac{s^J}{s}\left(W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L)\right)\right] + \phi(C_R - C_L)\right]. \tag{10.3}$$

A subset $\mathcal{L}$ of the groups are organized in lobbies. As in Section 8, $\lambda_{\mathcal{L}}$ denotes the organized fraction of the population. Lobby J maximizes the expected utility derived from economic policy, net the per capita cost of paying the contributions, namely

$$\left[p^R W^J(\boldsymbol{g}_R) + (1 - p^R) W^J(\boldsymbol{g}_L) - ((C_L^J)^2 + (C_R^J)^2)\right], \tag{10.4}$$

where C_L^J and C_R^J are the campaign contributions by lobby J to the parties, both constrained to be non-negative. Note that the cost of lobbying is taken to be a convex function of C^J, the last term on the right-hand side of Equation (10.4)[49]. In a richer model, this could reflect increasing marginal costs of enticing potential contributors with different willingness to give, where the lobby would naturally start by tapping those members of the group from whom collecting is easiest. Alternatively, if C represents contributions in kind, such as work in the campaign, the convexity may represent increasing disutility of effort. Whatever the interpretation, the total contributions received by party R are $C_R = \frac{1}{\mathcal{J}} \sum_{i \in \mathcal{L}} C_R^J$, and, similarly, for party L.

The timing of events is as follows: (i) Both parties simultaneously announce policy platforms. (ii) Having observed these announcements, all lobbies simultaneously set their campaign contributions. (iii) Elections are held. Stages (i) and (ii) are thus reversed relative to Section 8, where the lobbies instead moved first by setting contingent contribution schedules. The present timing assumption considerably simplifies the analysis and might also be more plausible. It portrays lobbying as an activity attempting to influence the electoral process, given the promises made by the parties. Note, however, that lobbying still influences policy formation, as parties anticipate how the lobbies will adapt their contributions to the parties' policy promises. Intuitively, each party wants to win the election; and one way of winning is to announce

[49] With linear cost functions for C^J, the reaction functions of the lobbies would not be continuous in the policy platforms in this set-up.

a platform appealing to the lobbies, and let the lobbies help garner electoral support by raising money or working for the party[50].

We are now ready to characterize the equilibrium. The electoral outcome at stage (iii) has already been discussed. Consider the optimization problem faced by the lobbies at stage (ii), for given policy platforms announced at stage (i). Maximization of Equation (10.4) with respect to C_R^J and C_L^J, subject to condition (10.3), yields[51]

$$\begin{aligned} C_R^J &= \text{Max}\left[0, \frac{\phi h}{\mathcal{J}}\left(W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L)\right)\right], \\ C_L^J &= -\text{Min}\left[0, \frac{\phi h}{\mathcal{J}}\left(W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L)\right)\right]. \end{aligned} \tag{10.5}$$

By Equation (10.5) each lobby campaigns only in favor of a single party, and does not campaign at all if the two parties announce identical platforms. This feature of the model is quite sensible – the lobbies want to influence the voters, not the parties – and it is consistent with some available evidence suggesting that lobbies seldom spend for both candidates in elections[52]. Summing this expression across all lobbies in $\mathcal{L}$, we get

$$C_R - C_L = \frac{\phi h}{\mathcal{J}^2} \sum_{J \in \mathcal{L}} \left[W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L)\right]. \tag{10.6}$$

That is, campaign spending goes to the party that is, on average, more successful in pleasing the lobbies.

Let us now turn to the party-optimization problem. Here, maximizing the vote share and the probability of winning amount to the same thing. By Equations (10.3), (10.2) and (10.6), party R's objective function can then be written

$$\text{Max} \frac{h}{\mathcal{J}} \left[\sum_J \frac{s^J}{s}\left[W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L)\right] + \gamma \sum_{J \in \mathcal{L}} \left[W^J(\boldsymbol{g}_R) - W^J(\boldsymbol{g}_L)\right]\right], \tag{10.7}$$

where $\gamma = h\phi^2/\mathcal{J} > 0$ is an extra weight on the lobbies' utility, related to how effective campaign spending is in influencing the voters: the more influential it is, the greater is the weight on the lobbies' utilities. Note the similarity with the assumed reduced-form objective of the government in the common-agency model in Section 8; in that case,

[50] Grossman and Helpman (1996) instead consider a set-up where the lobbies move first, and derive rather similar results.

[51] To derive Equation (10.5), note that by Equation (10.3) we have: $\frac{\partial p^R}{\partial C_R^J} = \frac{\phi h}{\mathcal{J}} = -\frac{\partial p^L}{\partial C_L^J}$; also recall that contributions are non-negative.

[52] For US evidence on this point, see Poole and Romer (1985).

the organized lobbies also get an additional weight in the objective of the policymaker. Thus, γ in the present model closely corresponds to $(1 - \eta)$ in Section 8.

By the same logic, party L solves an identical problem. Hence, like in Section 9, both parties announce the same policies: $\boldsymbol{g}_R = \boldsymbol{g}_L$, which then imply that equilibrium campaign spending is zero – cf. Equation (10.6)[53]. This does not mean that the presence of the lobbies is irrelevant; on the contrary: out of equilibrium, they do spend on the party who pleases them most, and this induces both parties to tilt public policy in their favor. Specifically, taking the first-order conditions of problem (10.7) and rewriting them, we can define the equilibrium allocation by the following expressions:

$$\begin{aligned} H_g(g^J) - 1 &= \frac{1}{s^J}\left[s - s^J + s\gamma\lambda_{\mathcal{L}}\right] \quad \text{if} \quad J \notin \mathcal{L}, \\ H_g(g^J) - 1 &= \frac{1}{s^J + \gamma s}\left[s - s^J - s\gamma(1 - \lambda_{\mathcal{L}})\right] \quad \text{if} \quad J \in \mathcal{L}. \end{aligned} \tag{10.8}$$

That is, g^J is overprovided, relative to the social optimum, if there are many swing voters in J (s^J is larger than s, the average of the other groups), precisely as in Section 9. If group J is organized as a lobby, there is also overprovision, and the lobbying effect is stronger, the higher is γ, i.e. the more effective are campaign contributions in influencing the voters. Also, a smaller fraction of lobbies among the groups, a smaller $\lambda_{\mathcal{L}}$, increases the overprovision for the lobbies, but decreases the underprovision for the unorganized groups, as in Section 8.

The model can easily be generalized to introduce other attributes of the voters. As noted above, Grossman and Helpman (1996) and Baron (1994) distinguish between *informed* and *uninformed* voters. The former are fully informed and completely unaffected by campaign contributions, like the voters in Section 9. The uninformed, on the other hand, are completely unaffected by economic policies, and their preferences only respond to campaign spending by the parties – namely their preferences are just given by the contributions term $\phi(C_R - C_L)$. Let groups also differ by the share of informed and uninformed voters, besides the density s^J, and let δ^J denote the share of informed voters in group J. Then, repeating the same steps as above, it can be shown that parameter δ^J influences the allocation, in the same way as s^J in expression (10.8). That is, groups with a larger share of informed voters are better treated by the parties, since they are more responsive to economic policies. Stated otherwise, voter mobility, one of the key determinants of the equilibrium allocation in the voting model, can either reflect a small weight given to ideology within the group (or small electoral turnout), or equivalently, a small share of uninformed voters.

This discussion naturally suggests two questions: How do voters obtain their information? And why are some voters informed while others are not? An obvious

[53] Grossman and Helpman (1996), with their different timing assumption, get a different result: in their model, there is non-convergence in party platforms, and equilibrium contributions are positive.

answer to the first question is that voters obtain their information from the media. Strömberg (1998) sets up a formal model of politics and the media to address the second question. He shows that the interaction between electoral competition (modeled as in Section 9) and competition between profit-maximizing media provides an answer to the second question. Optimal behavior by the media tends to bias the information – and hence also the policy outcome – towards groups that are attractive for advertisers.

To summarize, the model in this section provides a richer set of determinants of success in special-interest politics compared to the partial models in sections 8 and 9. But there are no surprises, and the results combine our earlier findings. As we shall see in the next two subsections, however, this is not always the outcome of interactions between different types of political activity.

10.2. Elections and legislative bargaining

To study the interaction between elections and legislation, we add an election stage at the beginning of the legislative-bargaining game above. In district-wide elections, forward-looking voters appoint a representative for the coming legislative session. As we shall see, this gives rise to strategic delegation, similar to that already encountered in the citizen-candidate model of Subsection 6.3. As in that section, we now assume that candidates are *outcome motivated*: they care about the policy enacted once in office, and different candidates have different views on what is the optimal policy. The modeling here follows quite closely the study by Chari, Jones and Marimon (1997).

Consider a four-stage game, where the last three stages are identical to the game in Section 7. In the first stage, every district simultaneously elects a representative by plurality rule. We assume that in each district, voters can choose among candidates with heterogeneous preferences for private versus public consumption. Specifically, a candidate of type α for district J has preferences

$$W^{J,\alpha} = c^J + \alpha H(g^J). \tag{10.9}$$

That is, candidates with high values of α care a great deal about publicly provided goods. Candidates are outcome motivated, in the sense that once elected, they act so as to maximize (10.9), and their type (ideology) is not an object of choice for the candidate himself. Candidates are thus characterized by their utility function (10.9), or, more compactly, by their preference parameter α.

For simplicity, we also make the following symmetry assumptions: (i) In all districts there is a continuum of candidates to choose from, with values of α in the same range $[\alpha^L, \alpha^U]$ for all districts. (ii) We continue to assume that *voters* are all identical within each district, and have preferences as in Equation (10.9), but with $\alpha = 1$. Adding voter heterogeneity – with voter preferences distributed over the same range $[\alpha^L, \alpha^U]$ as candidates – is straightforward and does not change the results. (iii) All districts have the same size, namely $\frac{N^J}{N} = \frac{1}{J}$ for all J. (iv) The default allocation is symmetric,

namely $\overline{g}^J = 0$ for all J, implying $\overline{\tau} = 0$. (v) Every representative has the same probability, $\frac{1}{\mathcal{J}}$, to be picked as the agenda setter.

Again we look for a Subgame perfect Nash equilibrium. Consider first the legislative-bargaining stages. By (iii), (iv) and the results in Section 7, it is easily shown that the chosen agenda setter will pick the $\frac{\mathcal{J}-1}{2}$ representatives with the highest values of α as members of the majority coalition, $\mathcal{M}$. The reason is that they are easiest to please, because they value public consumption a lot – i.e., their incentive constraints (7.1) are the easiest to relax. At the elections stage, voters realize this. Recall that voters in district J get compensated by some public goods for the taxes they pay only if their candidate is part of the majority, whereas they get no compensation if their candidate finds himself in the opposition. Hence, all districts have an incentive to elect a candidate with a value of α higher than that of the other districts, since that would make them part of the majority with certainty. This pushes all districts to a corner: under a mild condition on preferences, all districts elect the most spendthrift candidate, type α^U, in equilibrium. With this constellation of representatives, the voters in each district have a fifty–fifty chance of being included in the winning coalition. If any district appointed a "smaller spender" – a candidate with a lower α – this chance would drop to zero, thus bringing about a discontinuous expected welfare loss[54, 55].

Thus, we have another instance of strategic delegation: voters in each district elect a big spender. The reason is that unless they act in this way, they are left in the opposition. Clearly, this voting equilibrium makes the allocation more biased towards overspending for the agenda setter – since she also has a high α, on top of her better bargaining power – and diminishes the differences between districts inside and outside the majority.

As Chari, Jones and Marimon (1997) point out, this equilibrium is broadly consistent with opinions often expressed by American voters. Typically, they are quite disconcerted with the composition and actions of Congress as a whole but, at the same time, pleased with their own representative; the strong incumbency advantage

[54] Some conditions are needed to insure that this is an equilibrium, since electing a spendthrift candidate, the voters might also incur a cost: in the event that he is appointed agenda setter, a spendthrift ends up spending more than is optimal for his voters. This (expected) cost thus needs to be sufficiently smaller than the benefit, due to a discretely higher probability of being included in the majority. With a large enough number of districts, the probability of becoming agenda setter is sufficiently small, and this condition is satisfied.

[55] The model could be extended to an entry stage, where candidates sort themselves out as in the citizen-candidate model of Section 5. Suppose that voters too are heterogeneous and have the same preferences as the candidates, (10.9). Applying Proposition 2 (and Corollary 1) in Besley and Coate (1997), this equilibrium would, in fact, be sustainable in an extended "citizen-candidate" model with an initial entry stage, where every voter in each district could enter as a candidate, at a cost. The candidate with α^U optimally running and winning as an (unopposed) candidate in each district would be an equilibrium, if the entry cost was low enough and the default outcome bad enough (g^J valuable enough). See Coate (1997) for a full-fledged analysis of legislative bargaining and elections in a citizen-candidate model.

of serving legislators in congressional elections also bears testimony of this. In the equilibrium studied, voters in any district J would indeed have a higher expected utility if all other districts had representatives with $\alpha < \alpha^U$, but the voters in J could maintain the identity of their own representative.

The model is obviously very stylized, but still teaches us a lesson: it is not enough to look at the apparent bargaining powers that different legislators derive from a particular set of legislative rules, as these powers are endogenously modified in the interaction with their principals, the voters. Introducing elections thus pushes the legislative bargaining solution towards a more extreme outcome and not towards a more balanced one, as might have been one's first guess. The same point will reappear, even more forcefully, in the next subsection.

Nevertheless, the model neglects important aspects of the interactions between elections and legislative bargaining. Specifically, there is no connection between the election outcome and the proposal rights in the legislature. In reality, the allocation of these proposal rights is determined by the party affiliation and the seniority of legislators, and can be revised by each elected congress. In a remarkable paper, McKelvey and Riezman (1991) study these aspects in a dynamic game involving infinitely repeated elections in multiple districts, where each newly elected congress can set its own seniority rules before engaging in legislative bargaining over a fixed budget. McKelvey and Riezman show that seniority rights in agenda setting and a strong electoral incumbency advantage of senior legislators jointly emerge as a stationary equilibrium outcome. Interestingly, the endogenous seniority rights apply only to the initial proposal. If proposal rights in multi-round bargaining were to be given in the order of decreasing seniority, senior legislators would be at a disadvantage in the legislative bargaining. As they would have higher continuation values in each legislative session, it would be more expensive to bring them into the majority, in the same way as the votes of low-α legislators are more expensive in the model of this subsection.

10.3. Lobbying and legislative bargaining

We now set voters aside and consider how influence activities by interest groups interact with legislative bargaining. Research on this topic is still very scant. One antecedent is Snyder (1991), who studies how lobbies interact with legislators in the context of a spatial voting model[56]. A central insight is that lobbies will focus their contributions on "swing legislators", i.e., those who are indifferent between a proposal favorable to the lobbies and the status quo. Our analysis here draws on Helpman and Persson (1998).

[56] Another antecedent is Groseclose and Snyder (1996) who study a game where two lobbies buy votes from legislators about to decide on a public project. Interestingly, they show that when votes are bought sequentially, the prediction of a minimum winning coalition may fail.

With a structural model of government decision making, in place of a single policymaker, we must now take a stance on who lobbies whom. We restrict each interest group to make contributions only to a single congressman, "their own". This kind of fixed association is arbitrary but has some empirical support: campaign contributions in the USA tend to go to representatives from the same district as the donor, or to a member of the committee holding jurisdiction of regulation or grants applying to the donor group. For Europe there is much less systematic information about political contributions, but in some countries, there are very tight relations between interest groups, like trade unions and agricultural lobbies, and specific political parties[57].

Legislators still play the same legislative bargaining game. We retain symmetry assumptions (iii) and (iv) of the previous subsection. In addition, we also abstract from asymmetries in the organization across groups and assume that all groups are organized in lobbies: $|\mathcal{L}| = \mathcal{J}$ in the notation of Section 8. The policy game is as in Section 7, but with an additional contributions stage. The timing is as follows. First, Nature selects a legislator, $J = a$, to be the agenda setter. Then contribution schedules are simultaneously announced by the lobbies and observed by all legislators[58]. Finally, the agenda setter formulates a take-it-or-leave-it proposal, and the legislature votes on this. If the proposal is defeated, the default policy is as in the previous subsection: $\overline{\boldsymbol{g}} = 0 = \overline{\tau}$. We assume that legislators only care about the contributions they get.

Group J presents its congressional representative with a truthful contribution schedule, which offers

$$C^J(\boldsymbol{g}) = \begin{cases} \text{Max}[W^J(\boldsymbol{g}) - b^J, 0] & \text{if } g \text{ is supported by } J, \\ 0 & \text{otherwise,} \end{cases} \tag{10.10}$$

where the zero contribution if a policy $\boldsymbol{g}$ is not supported by legislator J can be shown to be an optimal strategy[59]. As in Section 8, we can think of b^J as reservation utilities of group J. Representatives maximize the value of their contributions, and hence want these reservation values to be as low as possible. As in Section 8, interest groups maximize their utility net of their contributions. Thus, they want the reservation utilities in Equation (10.10) to be as high as possible.

[57] Mueller (1989, ch. 11)) gives references to the empirical literature on campaign contributions in the USA. See Liebert (1995) for a discussion of lobbying in European parliamentary democracies.

[58] With the opposite timing (contributions made first), it would be natural to assume that contributions were made contingent on the status of the legislator (agenda setter or not). The results would be identical to the case considered in the text.

[59] Helpman and Persson (1998) show that indeed equilibrium contributions pay zero in the event that a legislator does not support a proposal. They also relax the assumption that legislators only care about money and show that the qualitative results are not affected if legislators also care about the welfare of their district.

Consider first the agenda setter's problem, for given contribution schedules. She wants to maximize

$$C^a(\boldsymbol{g}) = \text{Max}\,[W^a(\boldsymbol{g}) - b^a, 0] = \text{Max}\left[H(g^a) + y - \frac{1}{\mathcal{J}}\left(\sum_I g^I\right) - b^a, 0\right], \qquad (10.11)$$

subject to the incentive compatibility constraints that legislators in $\mathcal{M}$ are better off than with the default outcome:

$$W^J(\boldsymbol{g}) - b^J = H(g^J) + y - \frac{1}{\mathcal{J}}\left(\sum_I g^I\right) - b^J \geqslant 0 \quad \text{for} \quad J \in \mathcal{M} \qquad (10.12)$$

(recall that contributions are 0 if the proposal is voted down). Again, a finds it optimal to collect a minimum winning coalition, i.e. to include only $\frac{\mathcal{J}-1}{2}$ additional members in $\mathcal{M}$. It is easily shown that $\text{Max}[W^a(\boldsymbol{g})]$ is decreasing in all b^J, $J \in \mathcal{M}$. The agenda setter wants to satisfy condition (10.12) with equality for all members of the majority, as this maximizes her own district's utility and, hence, the contribution to herself. Thus, she picks the representatives with the lowest values of b^J as her coalition partners, setting $g^J = 0$ for everyone else, as in Section 9.

Now let us return to the contribution stage, and consider the optimal contributions for group J, for $J \neq a$. Clearly, group J is better off if its representative is included in the majority, as long as that gives at least a tiny piece of public goods[60]. This sets up a fierce "Bertrand competition" among the interest groups. As only legislators with the lowest reservation utilities are included in $\mathcal{M}$, the only equilibrium has every group J setting its reservation utility at the lowest possible level, namely $b^J = y - \frac{1}{\mathcal{J}}\left(\sum_{I \in \mathcal{M}} g^I\right)$. Returning to the agenda-setter's problem in Equations (10.11)–(10.12), we then find that the optimal solution satisfies

$$H_g(g^a) = \frac{1}{\mathcal{J}}, \qquad g^J = 0, \quad \text{all } J \neq a. \qquad (10.13)$$

Group a implements this choice at the lowest cost, namely zero, by setting its reservation utility $b^a = H(g^a) + y - \frac{g^a}{\mathcal{J}}$.

A useful way of thinking about this equilibrium is to rely on the same intuition as in the previous subsection. Each interest group badly wants to avoid that its representative be left in the minority, so that it only pays taxes but receives no public good. To avoid this outcome, each group reduces its reservation utility, so as to make the vote of her representative cheaper to buy. As all interest groups have the same objective, this

[60] If the representative is not included in the majority, the utility of group J is $W^J(\boldsymbol{g} \mid J \notin \mathcal{M}) = y - \frac{1}{\mathcal{J}}\left(\sum_{I \in \mathcal{M}} g^I\right)$, whereas the utility when she is included is $W^J(\boldsymbol{g} \mid J \in \mathcal{M}) = H(g^J) + y - \frac{1}{\mathcal{J}}\left(\sum_{I \in \mathcal{M}} g^I\right)$.

competition drives equilibrium public goods down to zero for every district. Obviously, the district of the agenda setter is the beneficiary. The logic is similar to that in Dixit, Grossman and Helpman (1997), who study a general common-agency model, and show that competition between the interest groups allows the single government to implement its preferred solution. But here, the benefit goes to one powerful district, not to a semi-benevolent government.

Note that also in this case, politicians collect no contributions in equilibrium. Clearly, this does not provide a safe ground for concluding that influence activities are unimportant, as some commentators like Tullock (1988) have suggested. Note also that in equilibrium, every legislator is willing to vote for the proposal (at least they do not have any incentive to vote against it). Thus, despite the force of minimum winning coalitions outside of equilibrium, the equilibrium majority is more than minimal. The model is thus consistent with a stylized fact, underlying the literature on "universalism" in the US Congress, namely that distributive bills often pass with broad majorities. But the universalism literature has weak micropolitical underpinnings (it is hard to model as the outcome of an extensive-form game), and universalism is often accounted for by referring to a "norm of deference" ("you scratch my back and I scratch yours"). In our setting we could imagine a sequence of legislative sessions, where different representatives (approximating different committees) take turns as agenda setters. The outcome after these sessions would coincide with a universalist allocation, like the one in Weingast, Shepsle and Johnsen (1981).

Note also that the results obtained in this section are *not* a convex combination of the results in the "partial" models studied above. Specifically, the distribution of benefits is more skewed than in the legislative-bargaining model of Section 7, even though the lobbying model of Section 8 predicted a very even distribution of benefits (with all groups organized and symmetric as we have assumed in this section, the common-agency model predicts equal b^J for all J).

These results illustrate, with additional force, the general point made in the previous subsection: optimal private behavior alters the bargaining powers inherent in legislative procedures. Here, they amplify the misallocation of public goods by a legislature where agenda-setting powers are conferred upon individual members or committees. Naturally, the simple structure of this game gives rise to an extreme outcome. Real-world legislatures have introduced various safeguards against such extreme outcomes. Some of these have already been discussed in Section 7, and some others will be discussed in Part III. We thus want to emphasize the general logic more than the specific results.

10.4. Notes on the literature

Our model of the interaction between elections and lobbying in Subsection 10.1 draws on Baron (1994), Grossman and Helpman (1996) and Bennedsen (1998). Besley and Coate (2001) study lobbying and elections in a citizen-candidate model; Riezman and Wilson (1997) study legal redistributions or contributions in a setting where

policymakers compete for the support of different lobbies. An early contribution on the interaction between lobbying and elections is Austen-Smith (1987).

The interaction between elections and legislative behavior is naturally of first-order importance in political economics. There is not much formal work combining extensive-form legislative games with elections and rational voters, which might be due to the difficulty of these issues. Austen-Smith and Banks (1988) and Baron (1993) are among the few that have studied the interaction between voting and government formation in a three-party setting. McKelvey and Riezman (1991) study the interactions between voting and legislative bargaining and show how a seniority system may emerge endogenously in a sequence of congressional elections. Subsection 10.2 draws on Chari, Jones and Marimon (1997). Coate (1997) demonstrates that the strategic delegation equilibrium considered by these authors is consistent with endogenous entry in a citizen-candidate model.

Work on the interdependencies between lobbying and legislation, assuming rational behavior of interest groups and legislators, is even more scarce. Denzau and Munger (1986) study a reduced-form model where interest groups give contributions to legislators who choose effort on different legislative activities so as to maximize expected votes. Snyder (1991) studies a structural model of the interactions between lobbies and legislators. Groseclose and Snyder (1996) study a game where two lobbies buy votes from legislators who will take a decision on a public project. Subsection 10.3 draws on Helpman and Persson (1998). Bennedsen and Feldman (2000) consider the interaction between lobbies and legislators in a context of incomplete information, where lobbies may provide information as well as campaign contributions to legislators.

Part III. Comparative politics

We often take it as given that democratic countries are representative, rather than direct, democracies. Yet, at a deeper level, the rationale and the implications for the delegation of *control rights* to elected office holders, rather than the delegation of mere administration, are not well understood. We may broadly consider the underlying reasons for delegation to be costly acquisitions of information. Unless the preferences of the citizens and their elected leaders are completely aligned, however, delegation of political control rights creates a principal–agent problem between the voters and their elected representatives. To minimize the adverse consequences of this agency problem then becomes one important role of the constitution.

The principal–agent relationship between voters and representatives entails some special features not always present in the agency problems typically studied by economists. First, voters are constrained to offer implicit rewards, through reappointment at elections, rather than explicit monetary incentives. Second, unbiased enforcement of detailed political contracts between politicians and voters at large may be problematic or impossible. Politics is the source of supreme authority: a

constitutional court would lose its legitimacy if it had to rule on detailed policy issues, and the appointment of judges could be difficult if their deliberations had a direct impact on the electoral success of one or the other of the political contenders. Third, an explicit contract between political representatives and voters may be unfeasible due to the complexity of the issues and the number of parties involved. Whatever the reasons, in the real world we do not observe complete constitutional contracts between voters and representatives. Political constitutions are typical examples of *incomplete contracts*: they allocate control rights over policymaking to different individuals or groups, in the same way as incomplete contracts allocate control rights to different stakeholders – such as equity holders, debtors, and managers – in a firm [Hart (1995)]. The study of comparative politics then becomes an investigation of how government policy decisions are shaped by the specific assignment of the proposal, amendment, veto and gate-keeping rights by the political regime, as well as the specific assignments of appointment rights by the electoral rule.

In this last part of the chapter, we discuss how this incomplete-contract perspective can be applied to public finance, paying particular attention to the agency problems. We are not interested in finding the optimal allocation of control rights, but rather in understanding the consequences of alternative forms of incomplete contracts for policy choices. Comparative politics, that is, the comparison of alternative political constitutions, thus amounts to comparing the consequences of alternative allocations of control rights over policy decisions. Even though the language is somewhat different – and the precise rationale for the delegation is rarely spelled out – such an agency and incomplete-contract approach is really at the core of the public-choice tradition, stemming from the seminal work of Buchanan and Tullock (1962); again, Mueller (1989, 1997) surveys the earlier literature. As in Part II, the more recent contributions are more explicit in spelling out specific constitutional details, and have gone further in assuming rationality by political actors.

We introduce the political agency problem in Section 11: elected representatives choose the supply of public goods and taxation, but can divert resources from the voters at large. All voters are alike and always unanimous. Thus, throughout the section, we exclusively focus on the conflict of interest between politicians and voters at large. In Subsection 11.1, we adapt the arguments in Barro (1973) and Ferejohn (1986) showing that elections create incentives for office-seeking politicians to behave in the voters' interest. But elections are not the only means of preventing abuse of power. As the founding fathers of the US constitution understood long ago, appropriate allocation of control rights – appropriate checks and balances – are also important. Following Persson, Roland and Tabellini (1997), we illustrate how to analyze these fundamental issues in political theory in Subsection 11.2. The main result is that "separation of powers", that is a specific allocation of proposal and veto rights, reduces equilibrium rents captured by politicians. The reason is that separation of powers creates a conflict of interest between elected officials. This is exploited by the voters, in order to limit the abuse resulting from contract incompleteness or asymmetric information.

In Section 12, we continue to assume that electoral promises cannot be enforced, but we drop the assumption that voters are unanimous. Now, policy can redistribute among voters, and different groups of voters seek to exploit the policymaking process to their advantage. As in Section 11, politicians can extract rents, so that there is a conflict between voters at large and politicians whether tax revenue should finance public goods or be appropriated in the form of rents. Finally, there is conflict among the politicians competing for these rents. We thus have a complicated multi-principal, multi-agent problem with conflicts in three dimensions: between voters and politicians, among voters, and among politicians. We ask how the electoral rule shapes equilibrium policy, contrasting some features of proportional vs. majoritarian elections.

In Section 13 we study the same general policy problem. But we turn from the rules governing elections to another set of consitutional features, namely the rules for legislation embodied in the political regime. More specifically, we contrast some stylized constitutional features of presidential and parliamentary democracies, drawing on recent work by Diermeier and Feddersen (1998) and Persson, Roland and Tabellini (1998). The assignment of control rights over legislation differs across these two regimes, which leads to very different equilibrium policy choices.

The issues discussed in this last part suggest many interesting institution-design questions, both positive and normative. How should constitutions be designed? Why do we observe the existing constitutions in different countries in the Western world? Are the theoretical predictions associating constitutional form and policy choice consistent with cross-country and time-series evidence mentioned in the General Introduction? At the current state of knowledge we have few answers, let alone satisfactory answers, to such fundamental questions. But we add a few brief remarks on how one might think about them in the last subsection of the chapter.

11. Agency costs and checks and balances

To analyze agency problems between voters and politicians in as simple and stark a form as possible, we start out by disregarding all conflicts of interest between different groups of voters. Thus, there are no redistributive instruments and all voters share the same preferences, namely

$$u = c + H(g) = y - \tau + H(g).$$

The notation is as before, with g now denoting a public good benefiting all voters. Population size is N and the government collects total tax revenue $N\tau$, which is used in the production of the public good. This production process allows politicians in office to appropriate some rents or squander some resources, r. These rents benefit politicians at the voters' expense. We thus write the budget constraint

$$g = N\tau - r,$$

where the size of r is also a policy choice. For example, we may think of r as a direct diversion of resources for private gain, as non-cost-effective defense purchases

benefiting the office holder or his friends, or as a bridge in the wrong place for most voters, but in the right place for a small group of benefactors. Due to unmodeled transaction costs, politicians only appropriate a fraction γ $[0 < \gamma < 1]$ of the resources r diverted from the provision of the public good, the rest is wasted. The size of these transaction costs could be determined by the transparency of the policymaking process, or by other institutional features relating to the execution of the budget, but here we just treat it as a parameter. Finally, we restrict g, τ and r to be non-negative.

Clearly, without the agency problem, the optimal policy from the viewpoint of the voters would always set $r = 0$ and raise enough revenue to allow a provision of public goods fulfilling the Samuelson criterion:

$$NH_g(g) = 1. \tag{11.1}$$

We now ask how far elections can go in enforcing this allocation.

11.1. Electoral accountability

It is a natural idea that electoral competition would discipline office-motivated candidates and limit rent extraction [Wittman (1989)]. It turns out that this is only true under special circumstances, namely if: (a) binding electoral promises are feasible, and (b) the two candidates are identical and hence perfect substitutes in the eyes of the voters. If either of these assumptions fails, in equilibrium political rents are generally positive.

In particular, even if electoral promises are fully binding, efficiency breaks down if the two candidates are perceived as different over some dimension by the voters, and hence as imperfect substitutes. This result is derived by Persson and Tabellini (1999b), Polo (1998) and Svensson (1997a) in probabilistic-voting models similar to those of Section 9. Intuitively, in a probabilistic-voting model the voters trade off economic efficiency against ideological affinity in some unspecified policy dimension. A candidate who announces larger rents than his opponent is punished by the voters, but not at an infinite rate: his probability of victory does not discontinuously jump to zero. Therefore, both candidates can optimally choose to precommit to grab positive rents if elected. Models of this kind are extensively discussed in Persson and Tabellini (2000).

But the assumption that any promise by politicians can indeed be enforced cannot be literally correct. Elections or electoral competition disciplines politicians through some sort of reputation mechanism, not because of outright enforcement. A politician who blatantly abused political powers for his own benefit would certainly be punished by the voters by not being re-elected. But note the implication; "not being re-elected" means that voters look *backward*, not *forward*. That is, elections perform the role of a disciplining device once policy has been chosen, rather than selecting among alternative policies. Good policies are rewarded by re-appointment, bad policies are punished by refusing re-election. In this subsection we illustrate this important role

of elections. We start by reviewing the original insights of Barro (1973) and Ferejohn (1986) in a model of electoral accountability, and then we briefly discuss some related ideas.

To study retrospective voting in the one-period model presented above, assume the following timing of events: (i) Voters set a reservation utility for re-electing the incumbent (see below). (ii) The incumbent policymaker freely sets policy; policy choices are observed by everybody. (iii) Elections are held, with the voters choosing between the incumbent and an opponent.

The incumbent's objective is to maximize

$$\mathsf{E}(v^I) = \gamma r + p^I R. \tag{11.2}$$

This objective reflects the incumbent policymaker's full discretion over current rents, r. What is at stake at the election are future rents, R, which can be interpreted as the expected present value of holding office from the next period and onwards. Here, we treat R as an exogenous parameter and neglect intertemporal discounting. But in a full intertemporal setting, R would be determined by the model[61].

At the election stage, the voters perceive no differences between the incumbent and the opponent: the two candidates are identical in the eyes of the voters, except for their past histories. Moreover, we assume that voters coordinate on the same retrospective voting strategy, punishing the incumbent for bad behavior and rewarding her for good behavior. This voting strategy boils down to setting the re-election probability p_I as follows:

$$p_I = \begin{cases} 1 & \text{if} \quad W(g,r) \geqslant \varpi, \\ 0 & \text{otherwise,} \end{cases} \tag{11.3}$$

where $W(g,r) \equiv y - (g+r)/N + H(g)$ is the voters' indirect utility from the observed policy, and ϖ is their reservation utility; below, we discuss how ϖ is chosen.

The voting strategy in Equation (11.3) creates a trade-off for the incumbent. When setting policy at stage (ii), she really has two alternatives. One option is to please the voters, giving them a policy which is rewarded with re-election and the pay-off R. In this case, the incumbent obviously wants to satisfy voters in the cheapest possible way, which implies choosing an efficient policy and keeping any slack as rents γr for herself. The total pay-off is $v = \gamma r + R$. The other option is to ignore re-election altogether and instead myopically maximize her rents in the manner of a Leviathan policymaker. This implies maximum taxation ($\tau = y$), no public good provision ($g = 0$) and maximal

[61] Ferejohn (1986) embeds a related one-period game in an infinite-horizon setting with exogenous benefits from office. Persson, Roland and Tabellini (1997) endogenize the future benefits from office, R, as the expected present value of endogenous rents from office, r, in future periods.

rents ($r = \gamma Ny$). Therefore, the incentive constraint under which the incumbent finds it (weakly) optimal to please the voters is:

$$\gamma r + R \geqslant \gamma Ny.$$

Voters cannot enforce lower rents than implied by this incentive constraint, but they clearly want it to be satisfied with equality. The minimum rent voters must collectively give up, in order not to trigger a myopic diversion, is thus

$$r^* = \text{Max}\left[0, Ny - \frac{R}{\gamma}\right]. \tag{11.4}$$

From the government budget constraint, $g^* + r^* \leqslant Ny$ (the maximum τ is given by y). Hence, for g^* to be affordable in equilibrium, we need

$$g^* \leqslant \frac{R}{\gamma}, \tag{11.5}$$

a condition we assume to be satisfied. Under this condition, voters obtain the optimal level of public goods, but if the right-hand side of Equation (11.4) is positive they must give up some rents.

What are the implications of this model? According to Equation (11.4), higher intrinsic value of public office (higher R) or higher rent-extraction costs (lower γ) contribute to keeping equilibrium rents down. But rents are higher if the tax base is higher (y higher). This reflects the source of rents, namely the discretion resulting from contractual incompleteness. A larger available tax base makes this discretion more threatening and the voters must give up more rents. Ferejohn (1986) and Persson, Roland and Tabellini (1997) also consider a stochastic setting, where the voters' utility for a given policy is random – for instance, the cost of producing the public good, or its value, may vary with the state of nature. If the policymaker can observe the state of nature but the voters cannot, she can exploit this information advantage by grabbing more rents. The equilibrium now has a bang–bang property. If the state of nature is favorable and the voters are easy to please, the incumbent seeks re-election; in unfavorable states of nature, on the other hand, pleasing the voters is too costly and the incumbent grabs as much rents as possible, knowing that she will then be ousted by the voters.

A critical feature of this simple model of electoral accountability is that voters are ex-post indifferent between reappointing the incumbent or voting for the opponent. Since the two candidates are identical, the only reason not to re-appoint the incumbent is to punish his bad behavior, rather than selecting a better or more honest politician. This feature may be plausible in some circumstances, but elections might also create incentives for good performance through other channels. An incumbent politician may want to impress the voters with his performance ahead of the elections, because voters are imperfectly informed about his ability, honesty, or other individual determinants of

good performance in office, and he wants to appear better than his opponent. Persson and Tabellini (2000) extend a career-concerns model originally due to Holmström (1982), which has precisely this implication. The model has three central assumptions: (a) Good performance reflects the talent of the politician in office. (b) Talent is a lasting attribute: a politician who is talented today has a high probability of also being talented tomorrow. (c) Talent is unobservable and unknown to voters (and in some versions of the model also to the politician himself, though this is less crucial). These three assumptions imply that an incumbent politician abstains from grabbing (unobserved) rents ahead of the elections, so as to increase his chances of appearing talented in the voters' eyes, and hence winning their re-appointment. Elections again act as a disciplining device against abuse of power, though for a different reason than in the model above.

A number of papers [reviewed in Persson and Tabellini (1999a) and Persson and Tabellini (2000)] have used this model to discuss electoral policy cycles. The incentives for good performance are particularly strong in the proximity of elections, and this can induce policy cycles, with better performance just ahead of the elections. In such papers, sharp electoral incentives can either be good or bad for the voters. They are good in that they deter incumbent politicians from abusing power and grabbing rents, as here. But they can also be bad for the voters, if an incumbent politician can create distortions to induce the *appearance* of good performance just ahead of the elections (for instance by over-spending in public goods and financing them with off-budget items, or by temporarily boosting growth with expansionary aggregate-demand policies).

11.2. Separation of powers

Once we begin to ask how to discipline opportunistic politicians, it is natural to consider other features of political institutions serving this purpose. These are old questions: ideas about the importance of constitutional checks and balances to prevent the abuse of political powers go back at least to Montesquieu and Locke, and played an important role in the federalist debate preceding the adoption of the US constitution. All political constitutions of the Western world to some degree incorporate the separation-of-powers principle. In this subsection, we show how a specific allocation of proposal and veto powers across different office holders may indeed make politicians more accountable to the voters. The result, due to Persson, Roland and Tabellini (1997), is similar to that already discussed in Subsection 7.2 with regard to overall spending. Here we adapt it to the example of this section, showing how sequential decision making and separation of powers might reduce equilibrium rents.

There are two political offices, the holders of which are simultaneously subject to re-election. We can consider these offices in different ways: as two legislative chambers, or as the executive and the legislative branch of government. In line with the latter interpretation, we label them the Executive, X and the Legislature, L. The general structure of the model is the same as in the past subsection. But the voters now

choose retrospective voting strategies for X and L separately. Total rents from office are split between the two office holders: $r^L + r^X = r$, and a specific policy decision must be made with regard to their allocation. Each incumbent office holder has an objective like (11.2),

$$\mathsf{E}(v^I) = \gamma r^I + p^I R^I,$$

except that I now refers to the office holders, $I = X, L$, rather than to competing parties, and R^I is the exogenous benefit of reappointment for the politician holding office I.

Consider a constitutional arrangement which, as in Subsection 7.2, imposes sequential decision making and separates sharp proposal powers over two policy dimensions. Specifically, consider the following game. (i) Voters choose a retrospective voting rule. (ii) The incumbent X proposes a tax rate τ. (iii) If the incumbent L approves, τ is implemented, otherwise a default tax rate $\tau = \bar{\tau} > 0$ is implemented. (iv) The incumbent L proposes a spending allocation $[g, r^L, r^X]$, subject to the tax rate from the prior stage: $g + r = N\tau$. (v) If X approves the proposal by L, it passes; if not, a default allocation $\bar{g} = \tau - \bar{r}^L - \bar{r}^X \geqslant 0$, $r^L = \bar{r}^L$, $r^X = \bar{r}^X$ is implemented. (vi) Voters observe g and τ. (vii) Elections are held where each incumbent runs against an identical opponent[62].

Sticking to the main interpretation, this arrangement thus implies a specific separation of political powers between the president and Congress in a presidential democracy. But it could also be interpreted as a separation of powers between the members of different standing committees in a congressional setting, or between different ministries in a parliamentary setting. Its effect is to strengthen the voters' ability to hold politicians accountable, thereby limiting equilibrium rents. If the two politicians have strong enough re-election incentives (in a sense specified below), voters can actually achieve the optimal solution with $r = r^L = r^X = 0$ and $g = g^*$.

To see why, suppose that voters have indeed adopted a demanding voting rule, conditioning the re-election of both incumbents on receiving first-best utility:

$$p^I = 1 \quad \text{for} \quad I = X, L \quad \text{iff} \quad W \geqslant y - \frac{g^*}{N} + H(g^*).$$

What are the incentives of the two office holders at the expenditure-decision stage (iv)–(v)? Their only chance of getting re-elected is if taxes have been set at the right level $\tau = g^*/N$ at the taxation stage (ii)–(iii). If so, L can either propose $r = 0$, $g = g^*$ and satisfy the voters, or else divert everything, setting $r = N\tau = g^*$. The former

[62] Note that the rents in the second-stage default, $\bar{r}^X$ and $\bar{r}^L$, are fixed numbers and do not depend on the first-stage decision. This is essential for the results stated below. As discussed later, separation of powers is only helpful under appropriate budgetary procedures, and our formulation of the default outcome is an essential part of these procedures.

choice gives L the pay-off R^L and X the pay-off R^X. Full diversion requires giving X at least $\gamma\bar{r}^X$ – as X knows she will not be re-elected, she requires at least the default pay-off not to veto a diversive proposal – making the net pay-off of L equal $\gamma(g^* - \bar{r}^X)$. Clearly, L prefers pleasing the voters if

$$g^* \leqslant \frac{R^L}{\gamma} + \bar{r}^X. \tag{11.6}$$

Does X have appropriate incentives to propose the right level of taxes at stage (ii)–(iii)? If she proposes $\tau = g^*$ and Equation (11.6) holds, then L will please the voters and X gets R^X. If she sets any other tax rate, L (who then cannot please the voters anymore) proposes maximal diversion and, according to the argument above, X nets $\bar{r}^X$. Thus, it is better for X to go along with the voters, if

$$\bar{r}^X \leqslant \frac{R^X}{\gamma}. \tag{11.7}$$

Finally, it is always better for L to accept such a proposal, unless the default level of taxes is too high[63].

If the value of office is high enough, in the sense that both (11.6) and (11.7) hold, the voters may thus credibly insist on the politicians delivering the unconstrained optimum. Adding these two conditions and using Equation (11.5), a necessary condition for full optimality is that the total value of office under separation of powers is at least as high as that without it: $R^L + R^X \geqslant R$. The agency problem of the previous section is thus completely eliminated, in the sense that equilibrium rents fall from r^* to zero.

Why does separation of powers strengthen accountability in this drastic way? The key is to remove from L, who controls the allocation of rents, any proposal powers over the size of the budget. The agent with proposal rights over taxes, X, is *not* a residual claimant on tax revenue, as L captures any additional rents created by higher taxes. This removes the conflict of interest between X and the voters. The only means whereby X can earn re-election is to set taxes at the level desired by the voters. A single office holder, instead, is always a full residual claimant on tax revenues; she can therefore threaten the voters with maximal diversion ($r = N\tau = Ny$); to avoid this Leviathan-like outcome, the voters must leave her some rents.

Note that separation of proposal powers is not enough, however, unless accompanied by appropriate checks and balances, also involving the allocation of amendment and veto rights. In this model, X only has veto rights, and is therefore nailed to its status-quo pay-off by the take-it-or-leave-it proposal by L in the last stage. This makes for a strong conflict of interest between X and L, that can be exploited by the voters. A more

[63] After a veto, leading to the tax rate $\bar{\tau}^S$, L will always make a diversive proposal at the next stage, giving her a pay-off of $\gamma(N\tau^S - \bar{r}^X)$. Thus a sufficient condition for L not to veto, given that the incentive-compatibility condition above holds, is that $\bar{\tau}^S \leqslant g^*/N$.

open bargaining procedure with amendment rights for X would make her a residual claimant on taxes and align the interest of the politicians against the interest of the voters. In this case, the benefit of separation of powers would be lost. In fact, separation of powers could even be detrimental for the voters, if it creates a common-pool problem among the two expected officials. This would happen if veto rights were removed and X and L could unilaterally determine how much to divert for themselves, r^X and r^L, with taxes or public consumption residually determined. Persson, Roland and Tabellini (1997) show that, in this case, equilibrium rents would be even higher than with a single policymaker. To put it differently, accountability can only work well if it is clear who is responsible for an observed abuse of power. The results in this section thus reinforce the general message anticipated in Subsection 7.2, about the importance of appropriate budgetary procedures and the virtues of multi-stage budgeting.

Separation of powers can also serve another purpose, namely to facilitate revelation of information to the voters. As discussed at the end of the previous subsection, private information enables politicians to earn informational rents. But separation of powers creates a conflict over the allocation of these rents, which helps the voters. In general, informational rents are earned by whoever has proposal powers over the allocation of spending, as he becomes the residual claimant on additional resources. This implies, however, that the other politician has no incentive to lie. In general, as shown by Persson, Roland and Tabellini (1997), the weak office holder's interests are aligned with those of the voters, who can then hold the powerful office holder accountable. Giving sharp proposal rights creating a conflict of interest between office holders enables the voters to eliminate all informational rents.

More generally, political accountability is more easily achieved if the constitution unambiguously allocates certain control rights to certain political offices. Naturally, this presupposes that separation of powers can be enforced, and that office holders do not re-allocate these control rights in other ways, for instance through collusive agreements.

11.3. Notes on the literature

The question of whether electoral competition induces opportunistic politicians to pursue efficient policies is an old one. The optimistic "Chicago school" is well represented by Stigler (1972) and Wittman (1989). For a more pessimistic view representing the "Virginia School", see Buchanan and Tullock (1962) and more of their followers.

Rents and electoral competition in a probabilistic-voting model were studied in Svensson (1997a), Polo (1998) and Persson and Tabellini (1999b). These papers also discuss various comparative-statics results, relating the size of equilibrium rents to political features such as the number of parties (Polo), the disagreement among voters (Svensson, Polo, Persson and Tabellini), and the electoral system (Persson and Tabellini). Svensson (1997b) presents empirical evidence that electoral accountability

works less well in politically polarized countries: such countries have higher government spending, but appear to have a less efficient public sector and lower growth. Mauro (1998) and Tanzi and Davoodi (1997) ask how corruption correlates with the composition of public spending in a large cross-section of countries.

Electoral accountability was first discussed in a principal–agent framework by Barro (1973) and then by Ferejohn (1986). Seabright (1996) stresses the incomplete-contract view and discusses electoral accountability by comparing different degrees of centralization in a federation. The model of Subsections 11.1 and 11.2 draws on Persson, Roland and Tabellini (1997), who emphasize the benefits of separation of powers. Separation of powers has also been discussed by Laffont and Martimort (1998), with regard to regulation by supervisory agencies. Laffont (1999) provides an excellent survey of the recent literature on collusion with politically motivated agencies.

Career concerns in a political context are studied by Persson and Tabellini (2000), who extend the seminal work of Holmström (1982). There is a large literature on electoral business cycles, based on variants of the career-concerns models with contributions by Lohmann (1996) and Rogoff (1990) among others. This literature is surveyed in Persson and Tabellini (1999a, 2000).

Finally, the literature on incomplete contracts is surveyed by Hart (1995) and Tirole (1999). Beyond the papers mentioned above, Aghion and Bolton (1998) discuss how to view constitutions as examples of incomplete contracts.

12. Electoral rules and public finance

So far, we have deliberately simplified the voters' task of holding their political agents accountable, by assuming that policy cannot redistribute between voters. We now relax this assumption, and allow economic policy also to redistribute among groups of voters. We continue to assume that binding electoral promises are not enforceable, and that elected politicians have the discretion to choose policy through legislative bargaining. Thus, we study policy choice in a genuine multiple-principal–multiple-agent setting. We now have conflict of interests running in three dimensions: between voters and politicians at large, over the size of aggregate rents; among voters, over the distribution of income; and among politicians, over the distribution of rents. How does the equilibrium provision of public goods to voters and rents to politicians interact with equilibrium redistribution across different groups of voters? And how do different electoral rules and the control rights laid down by different constitutions shape equilibrium policy in this richer setting? To try to answer these fundamental and difficult questions, we compare the equilibria under alternative stylized constitutions. Our goal is to capture the effects of alternative electoral rules on fiscal policy, and to compare stylized features of presidential-congressional and parliamentary systems.

In this section we retain the assumption of a single politician in office, and we compare alternative electoral rules for re-appointing him to office. Real-world electoral rules differ in several dimensions; two key dimensions are *district magnitude* and the

electoral formula. District magnitude is simply the number of legislators elected in a typical voting district. The electoral formula decides how vote shares are translated into seat shares. Here, the main distinction is between plurality rule – where only the n candidates obtaining the n highest vote shares, get the n seats awarded in a district – and proportional representation – where seat shares are proportional to vote shares.

While these concepts are analytically distinct, we find a strong correlation empirically; proportional representation tends to go hand in hand with large voting districts, whereas plurality rule tends to be combined with small districts.

The model presented in this section is constructed accordingly. Thus, we contrast proportional-representation elections with a single national district and plurality rule with many districts. Since a single incumbent is elected in office, we are literally comparing the electoral-college system for electing the US president (plurality rule with many districts), against a proportional-representation system with a single national district for electing the president. But hopefully the gist of this comparison also applies to legislative elections. The main insight of this Section is that plurality rule with several districts reduces the size of the minimum winning coalition, compared to proportional rule with a single national district. Following the contributions of Lizzeri and Persico (2001), Persson and Tabellini (1999b) and others, we show that this in turn shapes the incentives of politicians to provide public goods, as well as their incentives to grab rents. The section also briefly reviews other recent contributions on the effects of alternative electoral rules on fiscal policy more generally.

12.1. A simple policy problem

Consider the following version of the model of the previous section. There are N groups of voters indexed by J, all of size (mass) unity. Voters in group J have preferences

$$w^J = c^J + H(g) = y - \tau + f^J + H(g), \tag{12.1}$$

where the notation is the same as previously, except for f^J which denotes a lump-sum transfer to voters in group J. Even though voters only care about their net taxes (transfers), $\tau - f^J$, it is still important to distinguish between τ and f^J, because there are separate non-negativity constraints on these two instruments. As before, g denotes a general public good benefiting all voters.

A single incumbent politician holds office, and we continue to assume that he can appropriate rents, r. Hence, the government budget constraint is

$$N\tau = g + \sum_J f^J + r = g + f + r, \tag{12.2}$$

where f denotes aggregate transfers. All items in the government budget constraint must be non-negative.

Clearly, the social optimum for any symmetric (and strictly concave) social welfare function defined over the utility of voters, but not incorporating the rents to politicians, is to eliminate rents, setting $r = 0$, and to provide public goods in accordance with the Samuelson rule, $NH_g(g^*) = 1$. Net taxes $\tau - f^J$ should be equal across groups, implying $f^J = \frac{f}{N}$, even though optimal gross transfers are indeterminate. With a tiny bit of tax distortions, however, $f = 0$ becomes optimal.

But policy choice is delegated to a politician who maximizes the same objective function as in Equation (11.2), not to a benevolent social planner. Voters hold the incumbent accountable through retrospective voting strategies. The incumbent runs against an identical opponent.

We now want to know whether the equilibrium discussed in Subsection 11.1 can still be enforced in this richer setting, where we have added conflict among voters over redistributive transfers. We also want to know how the equilibrium policy depends on the features of the electoral rule.

12.2. Proportional representation with a single national district

Let us begin with proportional representation, where the incumbent runs for office in a single district, against an identical opponent.

A first result is that the benchmark equilibrium discussed in Subsection 11.1 breaks down. The reason is that the incumbent only needs to please a minimum winning coalition – i.e., a bare majority of the voters – to win re-election ($N/2$ voters, to keep the notation simple). Suppose all groups require the same level of utility as in the equilibrium of Subsection 11.1. If so, the incumbent can increase rents for himself by setting taxes at a maximum, $\tau = y$, reducing g somewhat, and offsetting all this by means of positive transfers f^J to $N/2$ voters to keep a majority satisfied. Since taxes fall on everyone while transfers are only given to half the voters, and since by Equation (11.1) the marginal utility of the public good is relatively small, he has the room to do this and strictly increase rents for himself. But some voters are hurt and do not reach their required reservation utility. Anticipating this outcome, these voters bid down their reservation utility, so as to be included in the minimum winning coalition.

An equilibrium must satisfy an additional optimality requirement. Let ϖ^J be the reservation utility chosen by group J. Then, in equilibrium, ϖ^J must be a *best response* to ϖ^I, for all $I \neq J$, taking into account what happens in the subsequent stages of the game. Thus, implicitly we are saying that voters within the group cooperate, by setting the same voting rule, but play Nash against all other groups. When this requirement is added, the equilibrium must have the following properties: (a) Voters must not be so demanding that the incumbent prefers to forego reappointment. (b) The equilibrium policy must be optimal for the incumbent, given that he only needs to please *a majority* of the voters to win reelection. (c) No group of voters can benefit from a unilateral change in their reservation utility, given what the other groups are asking.

Consider first properties (a) and (b). A policy vector satisfying these two properties can be computed as the solution of the problem of maximizing rents for the incumbent, subject to the government budget constraint (12.2), the usual non-negativity constraints, the upper bound on taxes ($\tau \leqslant y$), and the constraint that $N/2$ voters receive their reservation utility, namely:

$$y - \tau + f^J + H(g) \geqslant \varpi^J, \tag{12.3}$$

for some given ϖ^J. Property (c) implies that all voters must receive the same reservation utility. Because of the competition between different voter groups, in equilibrium, $f^J = 0$ for all J. Combining these requirements, we obtain that the equilibrium policy satisfies[64]:

$$\tau = y, \qquad NH_g(g) = 2, \qquad r = Ny - g. \tag{12.4}$$

Contrasting Equation (12.4) with (11.1), and (11.4), we immediately see that the presence of conflict among the voters makes them worse off compared to the equilibrium without any transfers. Note also that the incentive constraint is satisfied in equilibrium. The incumbent can now exploit the voters' conflict to his own benefit. As noted by Ferejohn (1986) in a related model, this reflects the contractual incompleteness at the core of this setting. As the opponent cannot promise that he will not play the disruptive game of pitting the groups of voters against each other, the voters are left at the incumbent's mercy. Ferejohn's (1986) model has no public good, only transfers and effort – the equivalent of (negative) rents – and equilibrium effort ends up being minimal. Here instead, there is an indivisible public good which puts an upper bound to the equilibrium rents. The indivisibility of the public good allows voters to set their reservation utility contingent on a measure of *aggregate* performance. Even though they do not act cooperatively, the public good provides an implicit coordination mechanism which helps the voters stop fighting each other and discipline the incumbent.

Note that, in equilibrium, we do not observe any redistributive transfers. But the mere possibility of resorting to this policy tool has a profound effect on equilibrium policy. This insight reminds us of the equilibria with lobbying and electoral competition discussed in Subsection 10.1, where no campaign contributions are observed in equilibrium, but the mere possibility of using them exerts a strong political influence.

12.3. Plurality rule with multiple districts

Suppose now that the election takes place in multiple districts according to plurality rule. The single incumbent in office now runs for re-election in $M < N$ electoral

[64] This equilibrium is computed as follows: maximize rents by choice of τ, g, $\{f^J\}$, subject to the constraints mentioned in the text and for given reservation utilities ϖ^J. This immediately gives the first two equations in (12.4). Then, add the requirement that in equilibrium, ϖ^J are the same for all voters. This implies $f^J = 0$, and hence the last equation in (12.4).

districts. To win reappointment, he now needs only half the votes in half the districts, as in the electoral college system for electing the US President. Districts are identical, and in each district there are N/M (groups of) voters. The equilibrium can be computed as in the previous subsection, except for one difference. To be reappointed, the incumbent needs to please $\frac{1}{2}\frac{N}{M}$ voters in $\frac{M}{2}$ districts. That is, he only needs to please $N/4$ voters. This means that, in computing the equilibrium, the incentive constraint (12.3) must now be satisfied for only $N/4$ voters, rather than for $N/2$ voters. By the same derivation as in the previous subsection, we can compute the equilibrium under majoritarian electoral rule. The expressions turn out to be identical to the expressions in (12.4), except for the equilibrium condition for the public good, which can now be written as

$$NH_g(g) = 4. \tag{12.5}$$

Public-good provision thus turns out to be even lower than with proportional elections. Furthermore, as $r = \tau - g = y - g$, rents are even higher.

To see the intuition, consider an electoral reform, from proportional to plurality rule. The reform allows the incumbent to make a profitable deviation from the previous equilibrium. By decreasing the supply of the public good from the point defined by $NH_g(g) = 2$ and raising the redistributive transfers for a quarter of the electorate, he can maintain a winning majority and still earn more rents. The operation reduces utility by $\frac{2\theta}{N}\Delta g$ for all voters, but releases Δg units of revenue. Compensating $\frac{N}{4}$ of the voters for the utility loss thus costs $\frac{2}{N}\Delta g\frac{N}{4} = \frac{\Delta g}{2}$, which leaves $\frac{\Delta g}{2}$ for additional rents. A deviation of this sort ceases to be profitable when public-good provision has reached the point given by Equation (12.5).

Under plurality rule, competition among the voters to be included in the winning coalition is even stiffer than under proportional representation, because the size of the minimum winning coalition has shrunk by half. The incumbent then takes advantage of this by pitting voters against each other to a greater extent. In equilibrium, the benefits of the public good are thus internalized for a smaller group of voters.

12.4. Discussion

From a positive point of view, the previous comparison suggests that proportional electoral systems lead to more public-good provision compared to majoritarian elections. The simple intuition is that proportional elections lead the incumbent to seek the support of a broader coalition of voters, compared to majoritarian elections. And public goods (or more generally programs that benefit a large coalition of beneficiaries, such as welfare-state programs) are an efficient instrument to reach that goal. Other papers based on different political models have obtained similar results. Lizzeri and Persico (2001) study a model of electoral competition à la Myerson (1993b). They show how the winner-takes-all property of plurality rule makes it more effective for political candidates to garner electoral support from small groups through narrowly targeted redistribution, at the expense of public-goods provision. Persson and Tabellini

(1999b) mix district size and the electoral formula as the model in this section. In equilibrium, proportional elections in a single national district induce more public-good provision, compared with plurality rule in several single member districts. The reason is that, under proportional elections, candidates need the support of half the voters in the population. They thus allocate public spending so as to benefit a large group of voters. With single member districts and plurality rule, instead, candidates only need to win in a few pivotal districts (they can neglect the districts where they are sure winners or sure losers). They thus have fewer incentives to provide public goods, preferring instead to target benefits towards smaller geographic groups of voters. Milesi-Ferretti, Perotti and Rostagno (2000) reach a similar conclusion, but in a very different model that combines legislative bargaining and strategic delegation, as in Subsection 10.3. They, as well as Persson and Tabellini (2001), find robust evidence in favor of the proposition that broad welfare-state programs (i.e., programs that are not narrowly targeted to geographic constituencies) tend to be larger in countries ruled by proportional elections, in line with the theoretical predictions.

The link between electoral rules and political rents has also been studied in the recent literature. In the model above we have stressed the idea that plurality rule in single member districts leads to more abuse of power because, by reducing the size of the minimum winning coalition, it exacerbates the conflict among voters. Myerson (1993a) gets the same prediction, based on different reasoning. His analysis suggests that single member districts and plurality rule raise barriers to entry in the political system, because fewer parties are typically represented in the legislature. If the parties represented differ in both ideology and intrinsic honesty, equilibrium rents reflect how much choice the voters have. With high barriers to entry (small electoral districts), the voters cannot easily punish dishonest politicians, because they would have to pay a high price in terms of ideological affinity. Hence the prediction that majoritarian elections are associated with more corruption and abuse of power by politicians. Ferejohn (1986) also obtains this prediction, in the electoral accountability model of Section 11. High barriers to entry make it more likely that a politician ousted from office re-enters in the next election. This reduces the deterrent effect of losing the elections, and leads to higher equilibrium rents.

Other features of majoritarian elections may have the opposite effect, however. In majoritarian elections, voters typically vote over single individuals. In proportional elections, instead, they vote on party lists. Persson and Tabellini (2000) argue that the latter, more indirect, chain of delegation – from voters to parties to candidates – dilutes the incentives to perform in the true interest of the voters, and thus raises equilibrium rents in proportional relative to majoritarian elections. Persson, Tabellini and Trebbi (2000) take these predictions to cross-country data, measuring rents by corruption as means used by the rankings of perceived abuse of political power compiled by Transparency International. They find that theoretical predictions are supported by the data: corruption tends to be higher the smaller is district magnitude (i.e. the higher are barriers to entry) and the greater is the fraction of legislators elected by voting on party lists (as opposed to individual candidates). But the second effect is quantitatively

more important and more robust, leading to the empirical conclusion that proportional electoral systems tend to be associated with more corruption.

12.5. Notes on the literature

A large and mostly empirical literature in political science compares different electoral rules and their effects within the political system. Some recent classics are Lijphart (1994, 1999), Taagepera and Shugart (1989) and Cox (1997). Laver and Shepsle (1990, 1996) and Schofield (1993) have studied cabinet formation in a spatial setting with many parties, but with no economic policy analysis.

Theoretical research by political scientists on the effect of different electoral rules within the political system is surveyed by Myerson (1995, 1999).

Research on the economic policy consequences of electoral rules is more scarce and more recent. It includes Myerson (1993b), Milesi-Ferretti, Perotti and Rostagno (2000), Morelli (1999), Lizzeri and Persico (2001), Persson and Tabellini (1999b, 2000), Austen-Smith (2000) and Rivière (1998).

Empirical evidence on the effect of different electoral rules on policy outcomes can be found in Milesi-Ferretti, Perotti and Rostagno (2000), Persson and Tabellini (1999b, 2001), and Persson, Tabellini and Trebbi (2000).

13. Political regimes and public finance

In Section 11, we illustrated the benefits of separation of power for holding politicians accountable. But what are the effects of alternative rules for legislative bargaining when there is also a conflict of interest between voters, and legislators must choose between a policy benefitting all voters (public-good provision) or some groups only (as with redistribution)? That is the question addressed in this section.

Specifically, we compare presidential-congressional regimes with parliamentary regimes. Several incumbents share office, and bargain over policies; we merge the analysis of Sections 11 and 12 with the legislative-bargaining approach of Section 7. In our stylized model of a presidential system, the responsibilities of politicians are more clearly defined, leading to more separation of powers. The stylized model of a parliamentary system, on the other hand, entails stronger incentives to form stable and broad coalitions. The presidential system therefore has more conflict among politicians as well as among voters. As we shall see, some earlier insights survive, but new results appear. It remains valid that separation of powers helps the voters control the agency problem: equilibrium rents are smaller in the presidential system. But more conflict also has costs, in that the presidential system supply less public goods and more targeted redistribution compared to the parliamentary system. The section draws on Diermeier and Feddersen (1998) and Persson, Roland and Tabellini (2000)

To make these points, we modify the model of Section 12, to allow for several politicians in office at the same time, as in Subsection 11.2. Specifically, suppose that

there are just three groups of voters ($N = 3$), all of size (mass) unity. Groups coincide with electoral districts, and each is represented by a single legislator, $l = 1, 2, 3$. Voters in district J have the same preferences as in (12.1). Aggregate rents r are now the sum of the rents appropriated by all legislators: $r = \sum_l r^l$. As before, each legislator maximizes the sum of his endogenous and (future) exogenous rents in office,

$$v^l = \gamma r^l + p^l R,$$

where p^l is the probability that legislator l is reappointed. For simplicity, the exogenous rents from office, R, are assumed to be equal for all incumbent law makers. Voters hold the incumbent law makers separately accountable in single member district elections. The incumbent legislator runs against an identical opponent in elections, which are held in each district after policy choices have been made.

We now discuss two different assignments of control rights over economic policies.

13.1. Congressional regime

A congressional-presidential system like that of the US has considerable separation of powers: different congressional committees hold proposal powers over legislation in different policy dimensions, and the President has veto power. To capture these features, we study a two-stage budget procedure where the proposal powers on taxes and on the allocation of spending are allocated to two different legislators. We thus abstract from the president and his veto powers, but these could be introduced without changing the thrust of the main results. We could further split the proposal power over spending further, giving each of the three legislators some agenda-setting privileges, without changing the main results.

The policy game studied has the following timing. (i) Two different agenda setters, a_τ and a_g, the "finance committee" and the "expenditure committee", are appointed among the three legislators. (ii) Voters set the cut-off utilities ϖ^J in their re-election rules optimally, conditional on the status of their legislator. (iii) a_τ proposes a tax rate τ. (iv) Congress votes: if approved by a majority, the tax proposal becomes law; if not, the default tax rate is $\bar{\tau} > 0$. (v) a_g proposes g, $\{f^J\}$ and $\{r^l\}$ subject to $3\tau \geqslant g + f + r$. (vi) Congress votes: if the proposal is rejected by a majority, the default allocation is $g = 0, f^J \equiv \tau - r^l \geqslant 0, r^l = \bar{r}$. (vii) Voters observe policy and elections are held.

As in subsection 11.2, there are thus two agenda setters. Policy decisions are made sequentially, first on the overall size of government and then on the allocation of spending. Not only are proposals sequential, but so are Congressional votes. Specifically, spending proposals in the second stage are constrained by the outcome of Congressional votes over tax revenues.

To understand the features of the equilibrium, we can draw on several results in previous sections[65]. In the last stage, the expenditure committee a_g just needs

[65] In the following, we just sketch the argument leading to the results. The reader is referred to Persson, Roland and Tabellini (2000) for a formal derivation in a similar (but more complex) set-up.

the support of one more legislator. Hence, as in Section 7, she seeks a minimum winning coalition in the legislature. Moreover, she seeks the support of the legislator who is "cheapest to buy", in the sense of demanding least for her constituency. Thus, voters in districts $J \neq a_g$ behave like the voters in Subsection 12.2: they become engaged in a "Bertrand competition" for the spoils allocated by a_g. Given that they pay taxes anyway, they are better off getting some transfers, however small, rather than zero. Hence, not to be excluded from the majority coalition, they reduce their reservation utilities until their demand for redistribution is driven to zero. Any equilibrium thus has $f^J = 0$ if $J \neq a_g$.

This leaves a_g free to please her voters, for all redistributive transfers go to her district ($f^J = f$ if $J = a_g$). The public good is then traded off against redistribution, one for one. This leads to severe underprovision of the public good, since only one third of the social benefits are internalized. Specifically, in equilibrium[66]:

$$g = \hat{g} \equiv H_g^{-1}(1).$$

What about equilibrium rents? As in Subsection 11.2, the maximum threat legislator a_g can impose on the voters is to go for the maximum diversion, $r^d = 3\tau$. Having bought the vote of one more law maker, she would be left with a pay-off of $\gamma(3\tau - \bar{r})$. Alternatively, she can satisfy the voters. Given that re-election is worth more than the default pay-off to the other legislators and that her proposal is consistent with the cut-off utilities demanded by the voters in the other districts [i.e., a condition like (11.7) holds], a_g is not obliged to pay any of the other legislators off with a positive r^J. Thus, pleasing the voters gives her the net pay-off of $\gamma r + R$. The incentive constraint on the minimum rents in stage (v) thus becomes

$$r \geqslant \text{Max}\left[3\tau - \frac{R}{\gamma} - \bar{r},\ 0\right]. \tag{13.1}$$

Finally, what are the incentives for the taxation committee a_τ and the voters in the corresponding district at the taxation stage (iii)? As *voters* in district a_τ do not receive any transfers, they would like τ to be as low as possible, consistent with $\hat{g}$ being financed. These interests are well aligned with those of *legislator* a_τ, for she is not a residual claimant on taxes, by our assumption – the sole residual claimants on additional revenue being legislator a_g or her voters. Voters in district a_τ will thus insist on the minimal tax rate, $\tau = \hat{g}/3$, implying that $r = f = 0$. Assuming as above that $R > \gamma\bar{r}$, it is optimal for a_g to go along with $r = 0$ in equilibrium. Similarly, voters

[66] In deriving this expression, we need to assume that the non-negativity constraint on transfers to the voters in $J = a^g$ is not binding. It can be shown that this assumption can be stated as $\hat{g} = H_g^{-1}(1) \leqslant \frac{R}{\gamma} + \bar{r}$.

in district $J \neq a_g, a_\tau$ have no reason to demand higher taxes from their legislator. The equilibrium is thus supported by voting rules with cut-off utilities:

$$\varpi^J = y + H(\widehat{g}) - \frac{\widehat{g}}{3}$$

for all voters[67].

We can summarize the properties of the congressional equilibrium as follows: First, taxes, rents and redistributive transfers are minimized: $\tau = \widehat{g}/3$ and $r = f = 0$. This follows from voters exploiting the separation-of-powers property of the congressional institution and from our assumption about the default outcome[68]. Second, public goods are severely underprovided: $H_g(\widehat{g}) = 1 > \frac{1}{3} = H_g(g^*)$. This is a direct consequence of the strong agenda-setting powers of a minority over the allocation of spending. Even with a larger amount of tax revenues, voters in the district who control the politician enjoying those powers would prefer to direct the available resources towards themselves, rather than sharing them with everyone, through more public-good provision. Anticipating this minoritarian orientation of redistributive transfers, voters in the district in charge of taxation keep tax revenues to the minimum necessary to provide the equilibrium amount of public goods.

13.2. Parliamentary regime

A central feature of the Presidential-congressional political regime described above is the non-stability of legislative coalitions: different coalitions are formed over different issues or at different points in time. This is at the core of the Bertrand-competition result, where legislators having control rights over the spending proposal can pit one group of voters against another. In parliamentary systems, on the other hand, disagreement within the majority in the legislature is a more serious business, since it can lead to a government crisis, through a defeat in a parliamentary vote of confidence. This creates an incentive for parliamentary coalitions to stick together – political scientists have labeled this feature of parliamentary systems "legislative cohesion". As a result, bargaining power is more evenly shared within the majority coalition. In our model, this is both good and bad for the voters. It is good, because it increases the equilibrium provision of public goods. It is bad because, by weakening separation of powers, it increases the equilibrium rents of politicians. We now formally derive these results in a simple extension of the previous model.

[67] We cannot rule out the existence of other equilibria with the same amount of r and g, but some positive redistribution to voters in $i = a^g$ and a higher tax rate.

[68] If the status-quo outcome $\bar{r}$ is positively related to τ raised in the taxation stage, it becomes harder to discipline the politicians and, as a result, the equilibrium has $r > 0$ [see Persson, Roland and Tabellini (2000)].

We continue to assume that two different legislators control the proposals on taxes and expenditures, respectively. No vote is taken, however, until both proposals have been made. It is therefore appropriate to identify these legislators with cabinet ministers and the proposal phase with the budget preparation inside the government. Both government coalition partners have veto power over the budget, and a veto triggers a government crisis. This assumption approximates having a vote of confidence attached to the government budget proposal. Obviously, this creates a strong incentive not to break up their coalition.

The new timing is: (i) Nature chooses a pair of representatives, who act as expenditure and finance ministers respectively: (a_g, a_τ). (ii) Voters set their reservation utilities conditional on the status of their legislators. (iii) The finance minister proposes a tax rate τ. (iv) The expenditure minister proposes expenditures $(g, \{f^J\}, \{r^J\})$, subject to the budget constraint and given the proposed tax rate. (v) Both members of government can veto the proposal. If neither of them does, the proposal passes and subsequently, elections are held. (v′) If at least one of them vetoes, the government breaks down and a default policy is implemented with $g' = \widehat{g}, f' = 0, r' = 3y - (\frac{R}{\gamma} + \bar{r}), r^{J\prime} = r'/3$, $\tau' = g' + r'$, and with re-election guaranteed for each legislator.

The default policy in (v′) may appear strange at first sight. Its pay-offs are designed to match the *expected* pay-offs for both voters and politicians after a government crisis in a more complex setting, where a government crisis leads to a new subgame. In this subgame, "a caretaker government" – a single legislator – is picked at random, voters reformulate their re-election rules, the caretaker legislator makes the entire budget proposal, and this is approved or not by the legislature [see Persson, Roland and Tabellini (2000)][69]. Not studying this subgame explicitly is, of course, a shortcut. But our assumption captures the essential feature, namely that the two government partners recognize that they have valuable agenda-setting powers inside the government and that a breakup is costly.

We now illustrate the equilibrium properties, referring the reader to Persson, Roland and Tabellini (2000) for a formal derivation. In this parliamentary regime, bargaining power is more equally shared among the coalition partners than in the Congressional regime. Hence, in this case, the final allocation splits welfare more equally among voters backing the majority coalition, as well as among their politicians. In particular, the equilibrium allocation of redistributive transfers and public goods must be jointly optimal for voters in the majority coalition. This generally leads to redistribution in favor of a majority, and the benefits of the public goods for the majority are internalized. That is, we have

$$\widehat{f}^J \geqslant 0, \quad J = a_\tau, a_g \quad \frac{1}{2} \leqslant H_g(\widehat{g}) < 1, \tag{13.2}$$

with $H_g(\widehat{g}) = \frac{1}{2}$ if $\widehat{f}^J > 0$ for both $J = a_\tau, a_g$.

[69] A richer model along the lines of Diermeier and Feddersen (1998) or Baron (1998) would have a new process of government formation following a crisis.

The equilibrium allocation is not unique, however. Since voters set their reservation utilities simultaneously, welfare can be split among them in many different ways. That is, bilateral monopoly now replaces Bertrand competition in the redistribution game between voters. All equilibria satisfy conditions (13.2). Hence, in all of these equilibria public-good provision is larger than in the Presidential system, and in most of them, redistributive transfers benefit a majority of voters.

On the other hand, equilibrium rents are higher than in the congressional regime, because separation of powers is no longer effective. As in Subsection 11.1, the government as a unified actor can impose the maximum threat of setting $\tau = y$ and $f = g = 0$ on the voters and forego re-election. To prevent this, voters must leave some rents to the governing coalition, at least to satisfy the joint incentive constraint: $r \geqslant 3y - 2R/\gamma$. Clearly, in equilibrium, the incentive constraint always binds, and equilibrium rents are $\hat{r} = 3y - 2R/\gamma$. This expression is almost identical to (11.4), except that here, the rents from office refer to two legislators rather than one. Aggregate rents are then split among legislators according to their bargaining power, which here reflects their veto rights[70].

Finally, voters in the majority now benefit from higher taxes, at the expense of the minority. Both legislators in the coalition are also pleased to go along with high taxes. Thus, in equilibrium, a_τ proposes $\hat{\tau} = y$ and a_g is pleased to accept it; voters in their districts are pleased as well[71].

The parliamentary equilibrium is thus different from the congressional equilibrium of the previous subsection in several respects. First, rents are unambiguously higher, as their mutual veto rights give both politicians in the coalition some bargaining power. Hence, they are both residual claimants of higher taxes, and voters can no longer exploit the conflict of interests between the legislators to their own benefit. Second, voters in the districts behind the stable majority are also pleased to support higher taxes, as the members of this majority jointly gain at the expense of the remaining minority. This majoritarian redistribution makes it less costly to provide public goods than in the congressional-presidential regime, however, and underprovision is less severe.

From a positive point of view, the analysis implies that parliamentary systems lead to a larger size of government compared to regimes with effective separation of powers and weaker incentives for legislative cohesion, such as presidential

[70] In particular, the finance minister will veto any proposal r^{a_τ} that does not give her at least as much as after a government crisis, namely $r'/3$. Note that politicians are re-elected in equilibrium as well as after the crisis.

[71] The parliamentary equilibrium is supported by the voting strategies

$$\varpi^{a_g} = f^{a_g} + H(\hat{g}), \qquad \varpi^{a_\tau} = 2\frac{R}{\gamma} - f^{a_g} - \hat{g} + H(\hat{g}).$$

Clearly, as f^{a_g} varies, so does the equilibrium utility of the two groups of voters, reflecting the multiplicity of equilibria.

systems. Persson and Tabellini (1999b, 2001) find strong empirical support for this prediction in a sample of 50–60 developed and developing democracies. Controlling for other variables, such as per capita income, the degree of openness of the economy, the age composition of the population, and other socio-economic variables, public spending is lower by about 10% of GDP in presidential regimes compared to parliamentary systems. Naturally, the theoretical models are very stylized, and it is a hard task to match the extensive forms of these games with observable institutional features. But the observed difference in spending between presidential and parliamentary systems is so large that the empirical result is likely to be robust to small errors in classifying regime types.

From a normative point of view, the analysis suggests a trade-off in institution design. In both political regimes, equilibrium policy differs from the social optimum: the institutional features that generate legislative cohesion also increase the rents to politicians, while separation of proposal powers induces legislative competition, and this, in turn, leads to more severe underprovision of public goods. Which distortion is worse depends on the circumstances. The parliamentary system appears better for the voters if the underprovision problem is large (because public goods are very valuable), while the presidential system dominates if the political agency problem is highly relevant (because politicians face small transaction costs in rent extraction, or the punishment from losing the next election is small, for instance due to barriers to entry in the political arena).

13.3. Concluding remarks

Sections 11–13 exemplify a number of interesting questions on how different allocations of political control rights shape equilibrium spending and taxation. A possible counterargument against such a research program in positive public finance is that it might involve a great deal of arbitrariness: "the possible combinations of control rights are infinite and you can prove anything with extensive-form game theory". While this may be a valid criticism against certain theories of industrial organization, we do not find it too damaging here. The reason is that constitutional rules are very well established, both legally and historically. Different democracies display a rich variation in the delegation of political control. A wealth of historical, descriptive and legal studies documenting these differences already exists. In other words, the rules – for proposing, amending or vetoing policy proposals, for forming or dissolving governments, or for electing political representatives – defining a particular extensive-form game need not be arbitrary, but can be given a solid empirical foundation.

Political scientists have done some analytical work, theoretical as well as empirical, on comparative politics. But that work is typically limited to consequences or correlations within the domain of the political system: certain electoral systems are found to be associated with a larger or smaller numbers of "effective parties", presidential systems are found to be more politically unstable than parliamentary systems, and so on. As already mentioned, there is also some work on economic

policy, for example on the correlation between different budget processes, different electoral systems and the propensity to run budget deficits. What is lacking is a systematic investigation of how commonly adopted constitutional arrangements shape fiscal policy choices. This kind of investigation sets a very interesting agenda for future empirical research, that some of the recent papers mentioned above have just started to explore [Milesi-Ferretti, Perotti and Rostagno (2000), Persson and Tabellini (1999b, 2001), Persson, Tabellini and Trebbi (2000)]. Aside from the general questions discussed at the end of the previous subsection, this agenda also includes other more specific questions. Does the recently adopted presidential line-item veto in the USA decrease or increase the equilibrium-policy favors granted to special interests? What kind of electoral reform could address the lack of political accountability that seems evident in countries like Japan, Italy or Belgium? Over what policy issues are referenda more likely to be desirable, and when might they be counter-productive? And so on.

Suppose we find mappings, by theoretical and empirical work, between political institutions and policy choices. What do we make of such results? Can we use them for normative recommendation of institutional reform, as hinted at the end of the previous section? Perhaps yes, perhaps no. One view is that this is futile, because constitutions, like policy choices, are endogenous and not subject to easy manipulation. In other branches of economics, like contract theory, information economics or corporate governance, the working assumption is often that observed institutions are efficient. Some researchers have also taken this view of political institutions.

We are sympathetic to the general idea of efficiency-oriented reform, but sceptical to its being used as an overall approach for understanding existing political institutions. Constitutional reforms are rare, due to the large transaction costs they involve. Unanticipated historical events may require new institutions, no matter how well-meaning were the constitutional framers. There is also a second argument. In some rare circumstances – like the US constitutional convention – constitutional reform may have taken place under a veil of ignorance about the future beneficiaries of certain rules. But reform is more often marginal, and reformers are often disinterested framers internalizing the desires of the average citizen. Rather, they tend to be active politicians who understand the conflicts of interests and participate in the political process after reform has taken place. In terms of our simple example in the previous section, suppose the agency problem dominates the underprovision of public goods from the point of view of the voters' welfare. Then, a constitutional assembly representing the voters at large would prefer a congressional system to limit political rents. But a constitutional choice made by politicians anticipating to be elected as representatives might instead prefer the parliamentary system. Thus, the agency problem re-appears at the level of constitutional choice.

13.4. Notes on the literature

Recent classics among political scientists, comparing different political regimes, include Bingham Powell (1982), Lijphart (1984) and Shugart and Carey (1992).

Tsebelis (1995) compares the role of veto rights in alternative political systems, while Huber (1996) studies the role of the motion of confidence in parliamentary systems.

The comparison between parliamentary and presidential-congressional systems in this section draws on Persson, Roland and Tabellini (1998, 2000) and Diermeier and Feddersen (1998); see also the work of Baron (1998) on legislative cohesion and government crisis. Breton (1991) also compares some features of parliamentary and congressional systems. Empirical evidence on size of government and public goods in presidential and parliamentary regimes is discussed in Persson and Tabellini (1999b, 2001).

A number of papers have investigated the empirical correlation between political institutions and budget deficits. See, in particular, Roubini and Sachs (1989), Grilli, Masciandaro and Tabellini (1991), Edwards and Tabellini (1994), Alesina and Perotti (1995) and Hallerberg and Von Hagen (1999).

The idea that economic institutions can be studied within the framework of contract theory, as optimal contractual arrangements, has been debated at length among economists, also contrasting complete and incomplete contracts. Coase (1960), Williamson (1985), Hart (1995), Tirole (1999) and Laffont (1999) express different views on this issue. Some researchers have also taken the view that political institutions can be studied as efficient arrangements. Wittman (1989, 1995) very explicitly applies this to the political system as a whole, while Krehbiel (1987), Gilligan and Krehbiel (1990) and Krehbiel (1991) take a similar approach in their information-based theory of the committee system. The idea that political institutions largely reflect the self-interest of politicians working within the system underlies another approach in the literature, which is common among rational-choice-oriented political scientists. These insights go back a long time, but are clearly exposed by Mayhew (1974), Fiorina (1977), and Weingast and Marshall (1988).

References

Aghion, P., and P. Bolton (1998), "Incomplete social contracts", Mimeo (University College London).

Aidt, T. (1998), "Political internalization of economic externalities and environmental policy", Journal of Public Economics 69:1–16.

Alesina, A. (1988), "Credibility and political convergence in a two-party system with rational voters", American Economic Review 78:796–805.

Alesina, A., and T. Bayoumi (1996), "The costs and benefits of fiscal rules: evidence from U.S. states", NBER Working Paper 5614 (NBER).

Alesina, A., and R. Perotti (1995), "The political economy of budget deficits", IMF Staff Papers 42:1–37.

Alesina, A., and D. Rodrik (1994), "Distributive politics and economic growth", Quarterly Journal of Economics 109:465–490.

Alesina, A., R. Hommes, R. Hausmann and E. Stein (1999), "Budget deficits and budget procedures in Latin America", Journal of Development Economics 59:233–253.

Arjona, R., and M. Pearson (2001), "Growth, inequality and social protection", OECD Working Paper (OECD).

Auerbach, A.J., and J.R. Hines Jr (2002), "Taxation and economic efficiency", in: A.J. Auerbach and M. Feldstein, eds., Handbook of Public Economics, Vol. 3 (Elsevier, Amsterdam) ch. 21, this volume.
Austen-Smith, D. (1987), "Interest groups, campaign contributions and probabilistic voting", Public Choice 54:123–139.
Austen-Smith, D. (1997), "Interest groups: money, information, and influence", in: D.C. Mueller, ed., Perspectives on Public Choice (Cambridge University Press, New York) pp. 296–321.
Austen-Smith, D. (2000), "Redistributing income under proportional representation", Journal of Political Economy 108(6):1235–1269.
Austen-Smith, D., and J.S. Banks (1988), "Elections, coalitions and legislative outcomes", American Political Science Review 82:405–422.
Austen-Smith, D., and J.R. Wright (1992), "Competitive lobbying for a legislator's vote", Social Choice and Welfare 9:229–257.
Azariadis, C., and V. Galasso (1997), "Fiscal constitutions and the determinacy of intergenerational tranfers", Working Paper 97-71 (Universidad Carlos III, Spain).
Baron, D. (1991), "Majoritarian incentives, pork barrel programs and procedural control", American Journal of Political Science 35:57–90.
Baron, D. (1993), "A theory of collective choice for government programs", Research Paper 1240 (Graduate School of Business, Stanford University).
Baron, D. (1994), "Electoral competition with informed and uninformed voters", American Political Science Review 88:33–47.
Baron, D. (1998), "Comparative dynamics of parliamentary governments", American Political Science Review 92:593–609.
Baron, D., and J. Ferejohn (1989), "Bargaining in legislatures", American Political Science Review 83:1181–1206.
Barro, R. (1973), "The control of politicians: an economic model", Public Choice 14:19–42.
Becker, G.S. (1983), "A theory of competition among pressure groups for political influence", Quarterly Journal of Economics 98:371–400.
Becker, G.S. (1985), "Public policies, pressure groups and deadweight costs", Journal of Public Economics 28:330–347.
Becker, G.S., and C.B. Mulligan (1998), "Deadweight costs and the size of government", Working Paper 6789 (National Bureau of Economic Research).
Benabou, R. (1996), "Inequality and growth", in: B. Bernanke and J. Rotemberg, eds., NBER Macroeconomics Annual (MIT Press, Cambridge) pp. 11–74,
Bennedsen, M. (1998), "Vote buying through resource allocation in government controlled enterprises", Mimeo (University of Copenhagen).
Bennedsen, M., and S. Feldman (2000), "Lobbying legislatures", Mimeo (University of Chicago).
Bernheim, B.D., and M.D. Whinston (1986), "Menu auctions, resource allocation, and economic influence", Quarterly Journal of Economics 101:1–31.
Bertola, G. (1998), "Microeconomic perspectives on aggregate labor markets", in: O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol. III (Elsevier, Amsterdam) pp. 2985–3028.
Besley, T., and S. Coate (1997), "An economic model of representative democracy", Quarterly Journal of Economics 112:85–114.
Besley, T., and S. Coate (1998a), "Centralized vs decentralized position of local public goods: a political economy analysis", Mimeo (London School of Economics).
Besley, T., and S. Coate (1998b), "Sources of inefficiency in a representative democracy: a dynamic analysis", American Economic Review 88:139–156.
Besley, T., and S. Coate (2001), "Lobbying and welfare in a representative democracy", Review of Economic Studies 68:67–82.
Bingham Powell Jr, G. (1982), Contemporary Democracies: Participation, Stability, and Violence (Harvard University Press, Cambridge, MA).

Blanchard, O.J., and J. Wolfers (2000), "The roles of shocks and institutions in the rise of European unemployment: the aggregative evidence," 1999 Harry Johnson Lecture, Economic Journal 100: C1–C33.

Boadway, R., and D.E. Wildasin (1989a), "Voting models of social security determination", in: B.A. Gustafsson and N.A. Klevmarken, eds., The Political Economy of Social Security (North-Holland, Amsterdam) pp. 29–50.

Boadway, R., and D.E. Wildasin (1989b), "A median voter model of social security", International Economic Review 30:307–328.

Boeri, T., A. Börsch-Supan and G. Tabellini (2001), "Would you like to shrink the welfare state? The opinions of European citizens", Economic Policy 32:7–50.

Bohn, H., and R.P. Inman (1996), "Balanced-budget rules and public deficits: evidence from the U.S. states", Carnegie Rochester Conference Series on Public Policy 45:13–76.

Boldrin, M., and A. Rustichini (2000), "Political equilibria with social security", Review of Economic Dynamics 3:41–78.

Bordignon, M., P. Manasse and G. Tabellini (2001), "Optimal intergenerational redistribution", American Economic Review 91:709–723.

Boylan, R.T. (1995), "An optimal auction perspective on lobbying", Mimeo (Washington University).

Breton, A. (1991), "The organization of competition in congressional and parliamentary governments", in: A. Breton et al., eds., The Competitive State (Kluwer Acedemic Publishers, Dordrecht) pp. 13–38.

Browning, E.K. (1975), "Why the social insurance budget is too large in a democracy?" Economic Inquiry 22:373–388.

Buchanan, J.M., and G. Tullock (1962), The Calculus of Consent. Logical Foundation of Constitutional Democracy (University of Michigan Press, Ann Arbor).

Buti, M., L.R. Pench and P. Sestito (1998), "European unemployment: contending theories and institutional complexities", Policy paper (European University Institute).

Chamley, C. (1986), "Optimal taxation of capital income in general equilibrium with infinite lives", Econometrica 54:607–622.

Chari, V.V., L.E. Jones and R. Marimon (1997), "The economics of split-ticket voting in representative democracies", American Economic Review 87:957–976.

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

Coate, S. (1997), "Distributive policy making as a source of inefficiency in representative democracy", Mimeo (University of Pennsylvania).

Cogan, J. (1994), "The dispersal of spending authority and federal budget deficits", in: J. Cogan, T. Murris and A. Schick, eds., The Budget Puzzle: Understanding Federal Spending (Stanford University Press).

Conesa, J.C., and D. Krueger (1999), "Social security reform with heterogeneous agents", Review of Economic Dynamics 2:757–795.

Cooley, T., and J. Soares (1999), "A positive theory of social security based on reputation", Journal of Political Economy 107:135–160.

Coughlin, P. (1992), Probabilistic Voting Theory (Cambridge University Press, Cambridge).

Coughlin, P., and S. Nitzan (1981), "Electoral outcomes with probabilistic voting and Nash social welfare maxima", Journal of Public Economics 15:113–121.

Cox, G. (1997), Making Votes Count: Strategic Coordination in the World's Electoral Systems (Cambridge University Press, New York).

Cox, G., and M. McCubbins (1986), "Electoral politics as a redistributive game", Journal of Politics 48:370–389.

Cukierman, A., and A. Meltzer (1989), "A political theory of government debt and deficits in a Neo-ricardian framework", American Economic Review 79:713–748.

Cukierman, A., and A. Meltzer (1991), "A political theory of progressive income taxation", in: A.H. Meltzer, A. Cukierman and S. Richard, eds., Political Economy (Oxford University Press, Oxford) pp. 76–108.

Daveri, F. (1998), "EFU after EMU?", Mimeo (IGIER).

Daveri, F., and G. Tabellini (1997), "Unemployment, growth and taxation in industrial countries", Mimeo, IGIER working paper 122 (IGIER).
Denzau, A., and M. Munger (1986), "Legislators and interest groups: how unorganized interests get represented", American Political Science Review 80:89–106.
Dewatripont, M., and E. Maskin (1995), "Credit and efficiency in centralized and decentralized Economies", The Review of Economic Studies 62(4):541–555.
Diermeier, D., and T. Feddersen (1998), "Cohesion in legislatures and the vote of confidence procedure", American Political Science Review 92:611–621.
Dixit, A.K. (1996a), The Making of Economic Policy: A Transaction-Cost Politics Perspective (MIT Press, Cambridge).
Dixit, A.K. (1996b), "Special interest politics and endogenous commodity taxation", Eastern Economic Journal 22:375–388.
Dixit, A.K., and J. Londregan (1996), "The determinants of success of special interests in redistributive politics", Journal of Politics 58:1132–1155.
Dixit, A.K., and J. Londregan (1998), "Ideology, tactics, and efficiency in redistributive politics", Quarterly Journal of Economics 113:497–529.
Dixit, A.K., G. Grossman and E. Helpman (1997), "Common agency and coordination: general theory and application to government policy making", Journal of Political Economy 105:752–769.
Downs, A. (1957), An Economic Theory of Democracy (Harper and Row, New York).
Drazen, A. (2000), Political Economy in Macroeconomics (Princeton University Press, Princeton, N.J).
Edwards, S., and G. Tabellini (1994), "Political instability, political weakness and inflation: an empirical analysis", in: C. Sims, ed., Advances in Economic Theory, Proceedings of the 1990 World Meetings of the Econometric Society (Cambridge University Press) pp. 355–376.
Enelow, J.M., and M.J. Hinich (1982), "Ideology, issues and the spatial theory of elections", American Political Science Review 76:493–501.
Feldstein, M., ed. (1998), "Privatizing social security", NBER Project Report (University of Chicago Press).
Ferejohn, J. (1986), "Incumbent performance and electoral control", Public Choice 50:5–26.
Ferejohn, J., and K. Krehbiel (1987), "The budget process and the size of the budget", American Journal of Political Sciences 31:296–320.
Ferejohn, J., M.P. Fiorina and R.D. McKelvey (1987), "Sophisticated voting and agenda independence in the distributive politics settings", American Journal of Political Sciences 31:169–194.
Fershtman, C., and K. Judd (1987), "Equilibrium incentives in oligopoly", American Economic Review 77:927–940.
Fiorina, M.P. (1977), Congress: Keystone of the Washington Establishment (Yale University Press, New Haven).
Fischer, S. (1980), "Dynamic inconsistency, cooperation, and the benevolent dissembling government", Journal of Political Economy 85:163–190.
Frey, B.S. (1983), Democratic Economic Policy – A Theoretical Introduction (Martin Robertson, Oxford).
Galasso, V., and J.I. Conde Ruiz (1999), "Positive arithmetic of the welfare state", Mimeo (Universidad Carlos III, Spain).
Gans, J.S., and M. Smart (1996), "Majority voting with single-crossing preferences", Journal of Public Economics 59:219–237.
Gilligan, T.V., and K. Krehbiel (1990), "Organization of informative committees by a rational legislature", American Journal of Political Science 34:531–564.
Grandmont, J.-M. (1978), "Intermediate preferences and the majority rule", Econometrica 46:317–330.
Grilli, V., D. Masciandaro and G. Tabellini (1991), "Political and monetary institutions and public financial policies in the industrial countries", Economic Policy 13:342–392.
Groseclose, T., and J. Snyder (1996), "Buying supermajorities", American Political Science Review 90:303–315.

Grossman, G., and E. Helpman (1994), "Protection for sale", American Economic Review 84:833–850.
Grossman, G., and E. Helpman (1995), "The politics of free-trade agreements", American Economic Review 85:667–690.
Grossman, G., and E. Helpman (1996), "Electoral competition and special interest politics", Review of Economic Studies 63:265–286.
Grossman, G., and E. Helpman (1998), "Intergenerational redistribution with short-lived governments", Economic Journal 108:1299–1329.
Grossman, G., and E. Helpman (2001), Special-Interest Politics (MIT Press, Cambridge, MA).
Hallerberg, M., and J. von Hagen (1999), "Electoral institutions, cabinet negotiations, and budget deficits within the European union", in: J. Poterba and J. von Hagen, eds., Fiscal Rules and Fiscal Performance (University of Chicago Press, Chicago) pp. 209–232.
Hart, O.D. (1995), Firms, Contracts and Financial Structure (Oxford University Press).
Hassler, J., and J.V. Rodriguez Mora (1999), "Employment turnover and unemployment insurance", Journal of Public Economics 73:55–83.
Hassler, J., J.V. Rodriguez Mora, K. Storesletten and F. Zilibotti (1998), "Equilibrium unemployment insurance", IIES Seminar Paper 665 (IIES).
Helpman, E., and T. Persson (1998), "Lobbying and legislative bargaining", Mimeo (Harvard University).
Hibbs, D.A. (1977), "Political parties and macroeconomic policy", American Political Science Review 71:1467–1497.
Holmström, B.R. (1982), "Managerial incentive problems – a dynamic perspective", in: Essays in Economics and Management in Honor of Lars Wahlbeck (Swedish School of Economics, Helsinki); Reprinted 1999 in Review of Economic Studies 66:169–182.
Huber, J.D. (1996), "The vote of confidence in parliamentary democracies", American Political Science Review 90:269–282.
Husted, T.A., and L.W. Kenny (1997), "The effect of the expansion of the voting franchise on the size of government", Journal of Political Economy 105:54–82.
Inman, R.P. (1987), "Markets, government and the 'new' political economy", in: A.J. Auerbach and M. Feldstein, eds., Handbook of Public Economics, Vol. 2 (North-Holland, Amsterdam) ch. 12.
Inman, R.P., and D.L. Rubinfeld (1997), "The political economy of federalism", in: D. Mueller, ed., Perspectives on Public Choice (Cambridge University Press, New York) pp. 73–105.
Inter-American Development Bank (1997), "Economic and social progress report 1997, Part III" (Inter-American Development Bank).
Judd, K.L. (1985), "Redistributive taxation in a simple foresight model", Journal of Public Economics 28:59–83.
Kau, J.B., and P.H. Rubin (1981), "The size of government", Public Choice 37:261–274.
King, M.A., and D. Fullerton, eds (1984), The Taxation of Income from Capital: A Comparative Study of the United States, United Kingdom, Sweden, and West Germany (University of Chicago Press, Chicago).
Kontopoulos, Y., and R. Perotti (1997), "Fragmented fiscal policy", Mimeo (Columbia University, New York).
Kontopoulos, Y., and R. Perotti (1999), "Government fragmentation and fiscal policy outcomes: evidence from the OECD countries", in: J. Poterba and J. von Hagen, eds., Fiscal Institutions and Fiscal Preference (Chicago University Press) pp. 81–102.
Kotlikoff, L.J., T. Persson and L.E.O. Svensson (1998), "Social contracts as assets: a possible solution to the time-consistency problem", American Economic Review 78:662–677.
Krehbiel, K. (1987), "Why are congressional committees powerful?", American Political Science Review 81:929–935.
Krehbiel, K. (1991), Information and Legislative Organization (University of Michigan Press, Ann Arbor).
Krusell, P., and V. Rios-Rull (1999), "On the size of the U.S. government: political economy in the neoclassical growth model", American Economic Review 89:1156–1181.

Krusell, P., V. Quadrini and V. Rios-Rull (1997), "Politico-economic equilibrium and economic growth", Journal of Economic Dynamics and Control 21:243–272.
La Porta, R., F. Lopez-de-Silane, A. Shleifer and R.W. Vishny (1999), "The quality of government", The Journal of Law, Economics and Organization 15:222–279.
Laffont, J.-J. (1999), Incentives and Political Economy, 1997 Clarendon Lectures (Oxford University Press).
Laffont, J.-J., and D. Martimort (1998), "Collusion and delegation", Rand Journal of Economics 30(2):232–262.
Laffont, J.-J., and J.-J. Tirole (1993), A Theory of Incentives in Procurement and Regulation (MIT Press, Cambridge, MA).
Lambertini, L., and C. Azariadis (1998), "The fiscal politics of big governments: do coalitions matter?", Mimeo (UCLA).
Laver, M., and K. Shepsle (1990), "Coalitions and cabinet government", American Political Science Review 81:873–890.
Laver, M., and K. Shepsle (1996), Making and Breaking Governments: Cabinets and Legislatures in Parliamentary Democracies (Cambridge University Press, New York).
Layard, R., and S. Nickell (1999), "Labour market institutions and economic performance", in: O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol. III (North-Holland, Amsterdam) pp. 3029–3084.
Ledyard, J.O. (1984), "The pure theory of large two-candidate elections", Public Choice 44:7–41.
Liebert, U. (1995), "Parliamentary lobby regimes", in: H. Döring, ed., Parliaments and Majority Rule in Western Europe (Campus-Verlag, Frankfurt).
Lijphart, A. (1984), Democracies (Yale University Press, New Haven).
Lijphart, A. (1994), Electoral Systems and Party Systems. A Study of Twenty-Seven Democracies 1945–1990 (Oxford University Press, Oxford).
Lijphart, A. (1999), Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries (Yale University Press).
Lindbeck, A., and J. Weibull (1987), "Balanced-budget redistribution as the outcome of political competition", Public Choice 52:273–297.
Lindert, P. (1994), "The rise of social spending, 1880–1930", Explorations in Economic History 31:1–36.
Lindert, P. (1996), "What limits social spending?" Explorations in Economic History 33:1–34.
Lizzeri, A., and N. Persico (2001), "The provision of public goods under alternative electoral incentives", American Economic Review 91:225–239.
Lockwood, B. (1998), "Distributive politics and the benefits of decentralization", Mimeo (University of Warwick, U.K.)
Loewy, M. (1988), "Equilibrium policy in an overlapping generations economy", Journal of Monetary Economics 22:485–500.
Lohmann, S. (1996), "Democracy and inflation", Mimeo (UCLA).
Lucas Jr, R.E. (1990), "Supply-side economics: an analytical review", Oxford Economic Papers 42: 293–316.
Mauro, P. (1998), "Corruption and the composition of government expenditure", Journal of Public Economics 69:263–280.
Mayhew, D.R. (1974), Congress: The Electoral Connection (Yale University Press, New Haven).
McCubbins, M., R. Noll and B.R. Weingast (1987), "Administrative procedures as instruments of political control", Journal of Law, Economics and Organization 3:243–279.
McKelvey, R.D. (1986), "Covering, dominance, and institution-free properties of social choice", American Journal of Political Science 30:283–314.
McKelvey, R.D., and R. Riezman (1991), "Seniority in legislatures", American Political Science Review 86:951–965.
Meltzer, A., and S. Richard (1981), "A rational theory of the size of government", Journal of Political Economy 89:914–927.

Meltzer, A., and S. Richard (1983), "Tests of a rational theory of the size of government", Public Choice 41:403–418.

Meltzer, A., and S. Richard (1985), "A positive theory of In-Kind Transfers and the negative income tax", Public Choice 47:231–265.

Mendoza, E., A. Razin and L. Tesar (1994), "Effective tax rates in macroeconomics: cross-country estimates of tax rates on factor incomes and consumption", Journal of Monetary Economics 34: 297–323.

Milesi-Ferretti, G.-M., R. Perotti and M. Rostagno (2000), "Electoral systems and the composition of public spending", Mimeo (Columbia University).

Morelli, M. (1999), "Equilibrium party structure and policy outcomes under different electoral systems", Mimeo (Iowa State University).

Mortensen, D.T., and C.A. Pissarides (1999), "New developments in models of search in the labor market", in: O.D. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol. III (North-Holland, Amsterdam) pp. 2567–2627.

Mueller, D. (1989), Public Choice II (Cambridge University Press, Cambridge).

Mueller, D. (1997), Perspectives on Public Choice (Cambridge University Press, New York).

Mulligan, C.B. (2001), "Economic limits on "rational" democratic redistribution", Mimeo (University of Chicago).

Mulligan, C.B., and X. Sala-i-Martin (1999a), "Social security in theory and practice, I: facts and political theories", Working Paper 7118 (National Bureau of Economic Research).

Mulligan, C.B., and X. Sala-i-Martin (1999b), "Social security in theory and practice, II: efficiency theories, narrative theories, and implications for reform", Working Paper 7119 (National Bureau of Economic Research).

Myerson, R. (1993a), "Effectiveness of electoral systems for reducing government corruption: a game theoretic analysis", Games and Economic Behavior 5:118–132.

Myerson, R. (1993b), "Incentives to cultivate favored minorities under alternative electoral systems", American Political Science Review 87:856–869.

Myerson, R. (1995), "Analysis of democratic institutions: structure, conduct and performance", Journal of Economic Perspectives 9(1):77–89.

Myerson, R. (1999), "Theoretical comparison of electoral systems", 1998 Schumpeter lecture, Berlin Congress of the European Economic Association; European Economic Review 43:671–697.

North, D.C. (1985), "The growth of goverment in the United States: an economic historian's perspective", Journal of Public Economics 28:383–399.

Olson, M. (1965), The Logic of Collective Action: Public Goods and the Theory of Groups (Harvard University Press, Cambridge, MA).

Osborne, M.J. (1995), "Spatial models of political competition under plurality rule: a survey of some explanations of the number of candidates and the positions they take", Canadian Journal of Economics 28:261–301.

Osborne, M.J., and A. Slivinsky (1996), "A model of political competition with citizen-candidates", Quarterly Journal of Economics 111:65–96.

Panizza, U. (1999), "On the determinants of fiscal centralization: theory and evidence", Journal of Public Economics 74:97–140.

Partridge, M. (1997), "Is inequality harmful for growth? Comment", American Economic Review 97:1019–1032.

Patton, C. (1978), "The politics of social security", in: M. Boskin, ed., The Crisis in Social Security (Institute for Contemporary Policy Studies, San Francisco) pp. 147–171.

Peltzman, S. (1980), "The growth of government", Journal of Law and Economics 19:211–240.

Perotti, R. (1996), "Growth, income distribution and democracy: what the data say", Journal of Economic Growth 1:149–188.

Persson, T. (1998), "Economic policy and special interest politics", Economic Journal 108:310–327.

Persson, T., and G. Tabellini (1990), Macroeconomic Policy, Credibility and Politics (Harwood Academic Publishers, London).
Persson, T., and G. Tabellini (1992), "The politics of 1992: fiscal policy and European integration", Review of Economic Studies 59:689–701.
Persson, T., and G. Tabellini (1994a), "Is inequality harmful for growth?" American Economic Review 84:600–621.
Persson, T., and G. Tabellini (1994b), "Representative democracy and capital taxation", Journal of Public Economics 55:53–70.
Persson, T., and G. Tabellini (1994c), "Does centralization increase the size of government?" European Economic Review 38:765–773.
Persson, T., and G. Tabellini (1995), "Double-edged incentives: institutions and policy coordination", in: G. Grossman and K. Rogoff, eds., Handbook of International Economics, Vol. III (North-Holland, Amsterdam) pp. 1973–2030.
Persson, T., and G. Tabellini (1996), "Federal fiscal constitutions: risk sharing and moral hazard", Econometrica 64:623–646.
Persson, T., and G. Tabellini (1999a), "Political economics and macroeconomic policy", in: J. Taylor and M. Woodford, eds., Handbook of Macroeconomics (North-Holland, Amsterdam) pp. 1397–1482.
Persson, T., and G. Tabellini (1999b), "The size and scope of government: comparative politics with rational politicians", 1998 Marshall Lecture; European Economic Review 43:699–735.
Persson, T., and G. Tabellini (2000), Political Economics: Explaining Economic Policy (MIT Press, Cambridge).
Persson, T., and G. Tabellini (2001), "Political institutions and policy outcomes: what are the stylized facts?", Mimeo (Stockholm University).
Persson, T., G. Roland and G. Tabellini (1997), "Separation of powers and political accountability", Quarterly Journal of Economics 112:1163–1202.
Persson, T., G. Roland and G. Tabellini (1998), "Towards micropolitical foundations of public finance", European Economic Review 42:685–694.
Persson, T., G. Roland and G. Tabellini (2000), "Comparative politics and public finance", Journal of Political Economy 108:1121–1161.
Persson, T., G. Tabellini and F. Trebbi (2000), "Electoral rules and corruption", Mimeo (Stockholm University).
Pissarides, C.A. (1990), Equilibrium Unemployment Theory (Blackwell, Oxford).
Polo, M. (1998), "Electoral competition and political rents", Mimeo (IGIER).
Poole, K., and T. Romer (1985), "Patterns of PAC contributions to the 1980 campaigns for the US House of Representatives", Public Choice 47:63–112.
Poterba, J.M. (1994), "State responses to fiscal crises: "natural experiments" for studying the effects of budgetary institutions", Journal of Political Economy 102:799–821.
Poterba, J.M., and J. von Hagen, eds (1999), Fiscal Rules and Fiscal Performance (University of Chicago Press, Chicago).
Potters, J., and F. van Winden (1992), "Lobbying and asymmetric information", Public Choice 74: 269–292.
Qian, Y., and G. Roland (1998), "Federalism and the soft budget constraint", American Economic Review 88:1143–1162.
Riezman, R., and J.D. Wilson (1997), "Political reform and trade policy", Journal of International Economics 42:67–90.
Riker, W.H. (1962), The Theory of Political Coalitions (Yale University Press, New Haven).
Rivière, A. (1998), "Strategic voting and electoral systems", Mimeo (ECARE).
Roberts, K. (1977), "Voting over income tax schedules", Journal of Public Economics 8:329–340.
Rodrik, D. (1995), "Political economy of trade policy", in: G. Grossman and K. Rogoff, eds., Handbook of International Economics, Vol. III (North-Holland, Amsterdam) pp. 1457–1494.

Rogoff, K. (1985), "The optimal degree of commitment of an intermediate monetary target", Quarterly Journal of Economics 100:1169–1190.
Rogoff, K. (1990), "Equilibrium political budget cycles", American Economic Review 80:21–36.
Romer, T. (1975), "Individual welfare, majority voting and the properties of a linear income tax", Journal of Public Economics 7:163–168.
Romer, T., and H. Rosenthal (1978), "Political resource allocation, controlled agendas and the status quo", Public Choice 33:27–43.
Romer, T., and H. Rosenthal (1979), "Bureaucrats versus voters: on the political economy of resource allocation by direct democracy", Quarterly Journal of Economics 93:563–587.
Rothstein, P. (1990), "Order restricted preferences and majority rule", Social Choice and Welfare 7:331–342.
Roubini, N., and J. Sachs (1989), "Political and economic determinants of budget deficits in the industrial democracies", European Economic Review 33:903–933.
Saint-Paul, G. (1993), "On the political economy of labor market flexibility", in: O.J. Blanchard and S. Fischer, eds., NBER Macroeconomics Annual (MIT Press, Cambridge) pp. 151–187.
Saint-Paul, G. (1996), "Exploring the political economy of labor market institutions", Economic Policy 23:265–315.
Schattschneider, E.E. (1935), Politics, Pressures and the Tariff (Prentice Hall, Englewood Cliffs, N.J.).
Schofield, N. (1993), "Party competition in a spatial model of coalition formation", in: W. Barnett, M.J. Hinch and N. Schofield, eds., Political Economy: Institutions, Competition and Representation (Cambridge University Press).
Scotchmer, S. (2002), "Local public goods and clubs", in: A.J. Auerbach and M. Feldstein, eds., Handbook of Public Economics, Vol. 4 (Elsevier, Amsterdam) forthcoming.
Seabright, P. (1996), "Accountability and decentralization in government: an incomplete contracts model", European Economic Review 40:61–89.
Shepsle, K. (1979), "Institutional arrangements and equilibrium in multidimensional voting models", American Journal of Political Science 23:27–59.
Shields, T.G., and R.K. Goidel (1997), "Participation rates, socioeconomic classes and Congressional elections", American Journal of Political Science 41:683–691.
Shugart, M.S., and J.M. Carey (1992), Presidents and Assemblies: Constitutional Design and Electoral Dynamics (Cambridge University Press, New York).
Siebert, H., ed. (1998), Redesigning Social Security (Instituet fuer Weltwirtschaft, Universitaet Kiel).
Snyder, J. (1991), "On buying legislatures", Economics and Politics 3:93–110.
Stigler, G.J. (1970), "Director's law of public income distribution", Journal of Law and Economics 13:1–10.
Stigler, G.J. (1971), "The theory of economic regulation", Bell Journal of Economics and Management Science 2:3–21.
Stigler, G.J. (1972), "Economic performance and political competition", Public Choice 13:91–106.
Strömberg, D. (1996), "Demography, voting, and local public expenditures: theory and evidence from Swedish municipalities", Mimeo (Princeton University).
Strömberg, D. (1998), "Mass-media competition, political competition, and public policy", Mimeo (Princeton University).
Stuart, C., and I. Hansson (1989), "Social security as trade among living generations", American Economic Review 79:1182–1195.
Svensson, J. (1997a), "The control of public policy: electoral competition, polarization and primary elections", Mimeo (The World Bank, Washington, D.C.).
Svensson, J. (1997b), "Accountability, polarization and growth: is the democracy better?" Mimeo (World Bank, Washington, DC).
Taagepera, R., and M.S. Shugart (1989), Seats & Votes. The Effects and Determinants of Electoral Systems, (Yale University Press, New Haven, Conn.).

Tabellini, G. (1991), "The politics of intergenerational redistribution", Journal of Political Economy 99:335–357.
Tabellini, G. (2000), "A positive theory of social security", Scandinavian Journal of Economics 102: 523–545.
Tanzi, V., and H. Davoodi (1997), "Corruption, public investment, and growth", IMF Working Paper 97/139 (IMF).
Tanzi, V., and L. Schuknecht (1995), "The growth of government and the reform of the state in industrial countries", IMF Working Paper 95/130 (IMF).
Tirole, J. (1999), "Incomplete contracts: Where do we stand?" Econometrica 67:741–781.
Tsebelis, G. (1995), "Decisionmaking in political systems: Presidentialism, parliamentarism, multicameralism, and multipartism", British Journal of Political Science 25:289–325.
Tullock, G. (1959), "Some problems of majority voting", Journal of Political Economy 67:571–579.
Tullock, G. (1988), "Future directions of rent-seeking research", in: K. Rowley, R. Tollison and G. Tullock, eds., The Political Economy of Rent-Seeking (Kluwer, Boston).
Verbon, H. (1988), The Evolution of Public Pension Schemes (Springer, Berlin).
Vickers, J. (1985), "Delegation and the theory of the firm", Economic Journal 95:138–147.
von Hagen, J. (1992), "Budget procedures and fiscal performance in the European communities", Economic Papers 96 (European Commission).
von Hagen, J. (1998), "Budgeting institutions for aggregate fiscal discipline", ZEI Policy Paper B98-01 (ZEI).
von Hagen, J., and I. Harden (1994), "National budget processes and fiscal performance", European Economy, Reports and Studies 3:311–418.
Weingast, B.R., and W. Marshall (1988), "The industrial organization of congress: or, why legislatures, like firms, are not organized as markets", Journal of Political Economy 96:132–163.
Weingast, B.R., K. Shepsle and C. Johnsen (1981), "The political economy of benefits and costs: a neoclassical approach to distributive politics", Journal of Political Economy 89:642–664.
Williamson, O.E. (1985), The Economic Institutions of Capitalism (Free Press, New York).
Wittman, D. (1977), "Candidates with policy preferences: a dynamic model", Journal of Economic Theory 14:180–189.
Wittman, D. (1983), "Candidate motivation: a synthesis of alternative theories", American Political Science Review 77:142–157.
Wittman, D. (1989), "Why democracies produce efficient results", Journal of Political Economy 97: 1395–1424.
Wittman, D. (1995), The Myth of Democratic Failure: Why Political Institutions are Efficient (University of Chicago Press, Chicago).
Wright, R. (1986), "The redistributive roles of unemployment insurance and the dynamics of voting", Journal of Public Economics 31:377–399.

Chapter 25

ECONOMIC ANALYSIS OF LAW

LOUIS KAPLOW and STEVEN SHAVELL
Harvard Law School and National Bureau of Economic Research

Contents

Handbook of Public Economics, Volume 3, Edited by A.J. Auerbach and M. Feldstein

Abstract

This is a survey of economic analysis of law, that is, of the emerging field under which the standard tools of microeconomics are employed to identify the effects of legal rules and their social desirability. Five basic subject areas are covered. The first is legal liability for harm. Here we discuss liability rules as incentives to reduce risk, issues of risk-bearing and insurance, and the costs of the liability system. Second, we consider property law, where we address the nature and justification of property rights, public property, the acquisition and transfer of property, externalities surrounding the use of property, and intellectual property. Third, we examine contract law, including the formation of contracts, their interpretation, and remedies for their breach. We focus on production contracts but also discuss other types, including donative contracts. Fourth, we treat the subject of civil litigation, that is, the bringing of lawsuits, and their settlement or disposition at trial. We also mention the appeals process, alternative dispute resolution, the provision of legal advice, and several additional topics relating to litigation. Fifth, we consider public enforcement of law, focusing on the level of law enforcement effort, the magnitude of sanctions, and other issues relevant to criminal law. Finally, we discuss criticisms that are commonly made by legal academics of economic analysis of law and offer concluding remarks.

Keywords

law and economics, liability, accident(s), tort(s), insurance, damage(s), causation, product liability, intentional tort(s), property, property right(s), externalities, regulation, firm(s), public property, takings, intellectual property, patent, copyright, trade secret, trademark, contract(s), disclosure, legal procedure, litigation, suit, settlement, trial, court(s), lawyer(s), judge(s), appeal(s), alternative dispute resolution, law enforcement, crime, plea bargaining, fine(s), imprisonment

JEL classification: K00, K10, K11, K12, K13, K14, K40, K41, K42, H23, L51

1. Introduction

Economic analysis of law seeks to answer two basic questions about legal rules. Namely, what are the effects of legal rules on the behavior of relevant actors? And are these effects of legal rules socially desirable? In answering these positive and normative questions, the approach employed in economic analysis of law is that used in economic analysis generally: the behavior of individuals and firms is described assuming that they are forward looking and rational, and the framework of welfare economics is adopted to assess the social desirability of outcomes.

The field of economic analysis of law may be said to have begun with Bentham (1789, 1827, 1830), who systematically examined how actors would behave in the face of legal incentives and who evaluated outcomes with respect to a clearly stated measure of social welfare (utilitarianism). Bentham's writings contain significant and extended analysis of criminal law and law enforcement, some analysis of property law, and a substantial treatment of the legal process. His work was left essentially undeveloped until the 1960s and early 1970s, when interest in economic analysis of law was stimulated by four important contributions: Coase's (1960) article on externalities and legal liability, Becker's (1968) article on crime and law enforcement, Calabresi's articles and culminating book (1970) on accident law, and R.A. Posner's (1972) general textbook on economic analysis of law and his establishment of the *Journal of Legal Studies*. As this survey will indicate, research in economic analysis of law has been active since the 1970s and is accelerating[1]. The field, however, is far from mature; one indication is the lack of empirical work on most topics.

Our focus here will be analytical, and we will cover five basic legal subjects[2]. The first three are the central areas of civil law. We begin with liability for accidents, which can be understood as addressing the problem of probabilistic externalities. Second, we discuss property law, which concerns the nature and justification of property rights, how they are acquired and transferred, how conflicts in the use of property are resolved, and related topics. Third, we examine contract law, including the formation of contracts, their interpretation, and remedies for their breach. The following section

[1] The field of law and economics is presented in a number of books, Cooter and Ulen (1997), Miceli (1997), Polinsky (1989), R.A. Posner (1998) and Shavell (forthcoming), and in two reference works, The New Palgrave Dictionary of Economics and the Law [Newman (1998)] and the Encyclopedia of Law and Economics [Bouckaert and De Geest (2000)]. Journals specializing in law and economics include the *Journal of Legal Studies,* the *Journal of Law and Economics,* the *Journal of Law, Economics, & Organization,* the *American Law and Economics Review,* and the *International Review of Law and Economics.* Also, professional organizations, including the American Law and Economics Association and the European Association of Law and Economics, are now well established.

[2] The sections on these subjects can be read largely independently of each other. Not treated in our survey are various, more particular areas of law than the five we have mentioned; omitted areas include antitrust law, corporate and securities law, bankruptcy and commercial law, banking law, international trade law, and tax law. Also excluded from this survey are problems addressed by the literatures on public choice and positive political theory.

concerns civil litigation, that is, the bringing of lawsuits by private actors to enforce their rights in the areas of law that we have just discussed. Next, we consider public enforcement of law, focusing on the level of law enforcement effort, the magnitude of sanctions, and other issues relevant to criminal law. Finally, we discuss criticisms that are commonly made by legal academics of economic analysis of law and offer concluding remarks.

2. Liability for accidents

Legal liability for accidents, which is governed by tort law, is a means by which society can reduce the risk of harm by threatening potential injurers with having to pay for the harms they cause. Liability is also frequently viewed as a device for compensating victims of harm, but we will emphasize that insurance can provide compensation more cheaply than the liability system. Thus, we will view the primary social function of the liability system as the provision of incentives to prevent harm.

There are two basic rules of liability. Under *strict liability,* an injurer must always pay for harm due to an accident that he causes. Under the *negligence rule,* an injurer must pay for harm caused only when he is found negligent, that is, only when his level of care was less than a standard of care chosen by the courts, often referred to as due care. (There are various versions of these rules that depend on whether victims' care was insufficient, as we will discuss below.) In fact, the negligence rule is the dominant form of liability; strict liability is reserved mainly for certain especially dangerous activities (such as the use of explosives).

Our discussion of liability begins by examining how liability rules create incentives to reduce risk. The allocation of risk and insurance will then be considered, and following that, the factor of administrative costs. Then we take up a number of important topics bearing on liability: the magnitude of liability (damages), causation, and the judgment-proof problem (assets insufficient to pay for harm). Finally, we consider the subjects of product liability and intentional torts[3].

2.1. Incentives

In order to focus on liability and incentives to reduce risk, we assume in this section that parties are risk neutral. Further, we suppose that there are two classes of parties, injurers and victims, and that they are strangers to one another, or at least are not in a contractual relationship. For example, injurers might be drivers and victims pedestrians, or injurers might be polluting firms and victims affected residents.

To begin with, we assume that accidents are unilateral in nature: only injurers can influence risks. Then we consider bilateral accidents, in which victims as well as

[3] A comprehensive economic treatment of accident law is contained in Shavell (1987a), which this section largely follows. See also Landes and Posner (1987a) and Calabresi (1970).

injurers affect risks. We also examine two types of action that parties can take that alter risk: first we consider their level of care (such as driving speed) and then their level of activity (number of miles driven).

2.1.1. *Unilateral accidents and the level of care*

Here we suppose that injurers alone can reduce risk by choosing a level of care. Let x be expenditures on care (or the money value of effort devoted to it) and $p(x)$ be the probability of an accident that causes harm h, where p is declining in x. Assume that the social objective is to minimize total expected costs, $x+p(x)h$, and let x^* denote the optimal x.

Under strict liability, injurers pay damages equal to h whenever an accident occurs, and they naturally bear the cost of care x. Thus, they minimize $x+p(x)h$; accordingly, they choose x^*.

Under the negligence rule, suppose that the due care level $\hat{x}$ is set equal to x^*, meaning that an injurer who causes harm will have to pay h if $x<x^*$ but will not have to pay anything if $x \geqslant x^*$. Then it can be shown that the injurer will choose x^*: clearly, the injurer will not choose x greater than x^*, for that will cost him more and he will escape liability by choosing merely x^*; and he will not choose $x<x^*$, for then he will be liable (in which case the analysis of strict liability shows that he would not choose $x<x^*$).

Thus, under both forms of liability, injurers are led to take optimal care. But note that under the negligence rule, courts need to be able to calculate optimal care x^* and to be able to observe actual care x, in addition to observing harm. In contrast, under strict liability courts do not need to do the former two; they only need to observe harm[4].

It should also be noted that, under the negligence rule with due care $\hat{x}$ equal to x^*, negligence would never actually be found, because injurers are induced to choose x^* and thus would be exonerated if they were sued after causing an accident. Findings of negligence may occur, however, under a variety of modifications of our assumptions. Courts might make errors in observing injurers' actual level of care so that an injurer whose true x is at least x^* might mistakenly be found negligent because his observed level of care is below x^*. Similarly, courts might err in calculating x^* and thus might set due care $\hat{x}$ above x^*. If so, an injurer who chooses x^* would be found negligent (even though care is accurately observed) because $\hat{x}$ exceeds x^*. As emphasized by Craswell and Calfee (1986), the chance of errors in the negligence determination leads injurers to choose incorrect levels of care; one possibility is that they would take excessive care in order to reduce the risk of being found negligent by mistake[5]. There

[4] Compare the discussion of corrective taxes versus regulation in Section 3.6.2.

[5] This might explain the phenomenon of "defensive medicine", on which see Danzon (1985) and, for empirical evidence, Kessler and McClellan (1996). Whether there is a tendency toward excessive

exist other explanations for findings of negligence as well, including that individuals may not know x^* and thus take too little care, the judgment-proof problem, which may lead individuals to choose to be negligent (see Section 2.6), and the inability of individuals to control their behavior perfectly at every moment or of firms to control their employees.

2.1.2. Bilateral accidents and levels of care

We now assume that victims also choose a level of care y, that the probability of an accident is $p(x,y)$ and is declining in both variables, that the social goal is to minimize $x+y+p(x,y)h$, and that the optimal levels of care x^* and y^* are positive[6].

Under strict liability, injurers' incentives are optimal conditional on victims' level of care, but victims have no incentive to take care because they are fully compensated for their losses. However, the usual strict liability rule that applies in bilateral situations is strict liability with a defense of *contributory negligence*, meaning that an injurer is liable for harm only if the victim's level of care was not negligent, that is, his level of care was at least his due care level $\hat{y}$. If victims' due care level is set by the courts to equal y^*, then it is a unique equilibrium for both injurers and victims to act optimally: victims can be shown to choose y^* in order to avoid having to bear their losses, and injurers will choose x^* since they will in fact be liable, as victims will not be negligent[7].

Under the negligence rule, optimal behavior, x^* and y^*, is also the unique equilibrium. Injurers can be shown to choose x^* to avoid being liable, and since victims will therefore bear their losses, they will choose y^*[8]. Two other variants of the negligence rule are negligence with the defense of contributory negligence (under which a negligent injurer is liable only if the victim is not negligent) and the

care depends upon the degree of legal error and on whether injurers who are found negligent are held responsible for all harm caused or only the incremental harm attributable to their negligence. On the latter, see Grady (1983) and Kahan (1989).

[6] In some early, less formal literature on accidents, for example, Calabresi (1970), reference is made to the notion of the "least-cost avoider", the party – injurer or victim – who can avoid an accident at the lower cost. The idea of a least-cost avoider relies on the assumption that each party can undertake a discrete amount of care that is independently sufficient to prevent an accident.

[7] That this equilibrium is unique follows from three observations: (1) Victims never have an incentive to take care y exceeding y^* (for once they take due care they will be compensated for their losses). (2) Victims will not choose y less than y^*, for if they do so, they will bear their own losses, injurers will take no care, and victims thus will minimize $y+p(0,y)h$. But $y+p(0,y)h = 0+y+p(0,y)h > x^*+y^*+p(x^*,y^*)h > y^*$, implying that victims must be better off choosing due care y^* than any $y<y^*$. (3) Because in equilibrium victims thus take due care, injurers choose x to minimize $x+p(x,y^*)h$, which is minimized at x^*.

[8] Uniqueness is demonstrated by the following: (1) Injurers will not take care exceeding x^*. (2) If injurers choose x less than x^*, victims will take no care, so injurers will minimize $x+p(x,0)h$, which exceeds $x^*+y^*+p(x^*,y^*)h$, which exceeds x^*. Thus, injurers are better off taking care of x^*. (3) Since injurers must choose x^* in equilibrium, victims will choose y^*.

comparative negligence rule (under which a negligent injurer is only partially liable if the victim is also negligent). These rules are also readily shown to induce optimal behavior in equilibrium.

Thus, all of the negligence rules, and strict liability with the defense of contributory negligence, support optimal levels of care x^* and y^* in equilibrium, assuming that due care levels are chosen optimally. Courts need to be able to calculate optimal care levels for at least one party under any of the rules, and in general this requires knowledge of the function $p(x,y)$. The main conclusions of this and the last section were first proved by Brown (1973)[9].

2.1.3. Unilateral accidents, level of care, and level of activity

Now let us reconsider unilateral accidents, allowing for injurers to choose their level of activity z, which is interpreted as the (continuously variable) number of times they engage in their activity (or if injurers are firms, the scale of their output). Let $b(z)$ be the benefit (or profit) from the activity, and assume the social object is to maximize $b(z)-z(x+p(x)h)$; here $x+p(x)h$ is assumed to be the cost of care and expected harm each time an injurer engages in his activity. Let x^* and z^* be optimal values. Note that x^* minimizes $x+p(x)h$, so x^* is as described above in Section 2.1.1, and that z^* is determined by $b'(z)=x^*+p(x^*)h$, which is to say, the marginal benefit from the activity equals the marginal social cost, comprising the sum of the cost of optimal care and expected accident losses (given optimal care).

Under strict liability, an injurer will choose both the level of care and the level of activity optimally, as his objective will be the same as the social objective, to maximize $b(z)-z(x+p(x)h)$, because damage payments equal h whenever harm occurs.

Under the negligence rule, an injurer will choose optimal care x^* as before, but his level of activity z will be socially excessive. In particular, because an injurer will escape liability by taking care of x^*, he will choose z to maximize $b(z)-zx^*$, so that z will satisfy $b'(z)=x^*$. The injurer's cost of raising his level of activity is only his cost of care x^*, which is less than the social cost, as it also includes $p(x^*)h$. The excessive level of activity under the negligence rule will be more important the larger is expected harm $p(x^*)h$ from the activity.

The failure of the negligence rule to control the level of activity arises because negligence is defined here (and for the most part in reality) in terms of care alone. A justification for this restriction in the definition of appropriate behavior is the difficulty courts would face in determining the optimal z^* and the actual z. Moreover, the problem with the activity level under the negligence rule is applicable to any aspect of behavior that would be difficult to regulate directly (including, for example, research and development activity). If, instead, courts were able to incorporate all aspects of

[9] Diamond (1974) proved closely related results shortly afterward. See also Green (1976), Emons (1990), and Emons and Sobel (1991), who focus on the case of heterogeneous injurers and victims.

behavior into the definition of negligence, the negligence rule would result in optimal behavior in all respects. (Note that the variable x in the original problem could be interpreted as a vector, with each element corresponding to a dimension of behavior.)

2.1.4. Bilateral accidents, levels of care, and levels of activity

If we consider levels of care and of activity for both injurers and victims, then none of the liability rules that we have considered leads to full optimality (assuming that activity levels are unobservable). As just explained, the negligence rule induces injurers to engage excessively in their activity. Similarly, strict liability with a defense of contributory negligence leads victims to engage excessively in their activity (the number of times they expose themselves to risk), as they do not bear their losses given that they take due care. The reason that full optimality cannot be achieved is in essence that injurers must bear accident losses to induce them to choose the right level of their activity, but this means that victims will not choose the optimal level of their activity, and conversely[10]. The distinction between levels of care and levels of activity was first emphasized in Shavell (1980c), where the results of this and the last section were shown.

2.1.5. Empirical evidence on the effect of liability on safety

Only a modest amount of empirical work has been undertaken on the effect of liability on accident risks. See Dewees, Duff and Trebilcock (1996) for a general survey of the literature that exists, and, among others, Devlin (1990), Landes (1982), and Sloan et al. (1994) on liability and auto accidents, Danzon (1985) and Kessler and McClellan (1996) on liability and adverse medical outcomes, and Higgins (1978), Priest (1988), and Viscusi (1991) on liability and product safety.

2.2. Risk-bearing and insurance

We consider next the implications of risk aversion and the role of insurance in the liability system, on which see Shavell (1982a). Several general points may be made.

First, the socially optimal resolution of the accident problem obviously now involves not only the reduction of losses from accidents, but also the protection of risk-averse parties against risk. Note that risk-bearing is relevant for two reasons: not only because potential victims may face the risk of accident losses, but also because potential injurers may face the risk of liability. The former risk can be mitigated through

[10] However, there exist ways to induce fully optimal behavior using tools other than conventional liability rules. For example, if injurers have to pay the state for harm caused and victims bear their own losses, both victims and injurers will choose levels of care and of activity optimally. On the possibility of such decoupling of what injurers pay and what victims receive, see note 107.

insurance that covers losses suffered in accidents, and the latter through liability insurance.

Second, because risk-averse individuals will tend to purchase insurance, the incentives associated with liability do not function in the direct way discussed in the last section, but instead are mediated by the terms of insurance policies. To illustrate, consider strict liability in the unilateral accident model with care alone allowed to vary, and assume that insurance is sold at actuarially fair rates. If injurers are risk averse and liability insurers can observe their levels of care, injurers will purchase full liability insurance coverage and their premiums will depend on their level of care; their premiums will equal $p(x)\,h$. Thus, injurers will want to minimize their costs of care plus premiums, or $x+p(x)\,h$, so they will choose the optimal level of care x^*. In this instance, liability insurance eliminates risk for injurers, and the situation reduces to the previously analyzed risk-neutral case.

If, however, liability insurers cannot observe levels of care, ownership of full coverage could create severe moral hazard, so would not be purchased. Instead, as we know from the theory of insurance, the typical amount of coverage purchased will be partial, for that leaves injurers with an incentive to reduce risk. In this case, therefore, the liability rule results in some direct incentive to take care because injurers are left bearing some risk after their purchase of liability insurance. But injurers' level of care will still tend to be less than first-best.

This last situation, in which liability insurance dilutes incentives, leads to our third point, concerning the question whether the sale of liability insurance is socially desirable. (We note that because of fears about incentives, the sale of liability insurance was delayed for decades in many countries and that it was not allowed in the former Soviet Union; further, in this country liability insurance is sometimes forbidden against certain types of liability, such as against punitive damages.) The answer to the question is that sale of liability insurance is socially desirable, at least in basic models of accidents and some variations of them. In the case just considered, the reason is evident. Injurers are made better off by the presence of liability insurance, as they choose to purchase it. Victims are indifferent to its purchase by injurers because victims are fully compensated under strict liability for any losses they sustain. In particular, it does not matter to victims that the likelihood of accident may rise due to the sale of liability insurance. This argument must be modified in other cases, such as when the damages injurers pay are less than harm because injurers are judgment proof. In that circumstance, the sale of liability insurance may not be socially desirable. See Section 2.6.

Fourth, consider how the comparison between strict liability and the negligence rule is affected by considerations of risk-bearing. It is true that the immediate effect of strict liability is to shift the risk of loss from victims to injurers, whereas the immediate effect of the negligence rule is to leave the risk on victims (injurers will tend to act non-negligently). However, the presence of insurance means that victims and injurers can substantially shield themselves from risk. Of course, as was just discussed, insurance coverage may be incomplete due to moral hazard; this makes risk-bearing of some

relevance to the comparison of liability rules, but which rule becomes more favorable is not obvious.

Finally, as we stated at the outset of Section 2, the presence of insurance implies that the liability system cannot be justified primarily as a means of compensating risk-averse victims against loss. Rather, the justification for the liability system must lie in significant part in the incentives that it creates to reduce risk. To amplify, although both the liability system and the insurance system can compensate victims, the liability system is much more expensive than the insurance system (see the next section)[11]. Accordingly, were there no social need to create incentives to reduce risk, it would be best to dispense with the liability system and to rely on insurance to accomplish compensation[12].

2.3. Administrative costs

2.3.1. Administrative costs of the liability system

The administrative costs of the liability system are the legal and other costs (notably the time of litigants) involved in bringing suit and resolving it through settlement or trial. These costs are substantial; a number of estimates suggest that, on average, administrative costs of a dollar or more are incurred for every dollar that a victim receives through the liability system. In contrast, the administrative cost of receiving a dollar through the insurance system is often below fifteen cents[13].

The factor of administrative costs affects the comparison between the forms of liability. On one hand, we would expect the volume of cases – and thus administrative costs – to be higher under strict liability than under the negligence rule. This is because, under strict liability, a victim can collect whether or not the injurer was at fault, whereas under the negligence rule fault must be established, so that in many cases of accident there will be no suit or, if there is a suit, it will be dropped after little has been spent[14]. On the other hand, given that there is a case, we would anticipate administrative costs to be higher under the negligence rule than under strict liability, because under the

[11] Also, victim compensation through liability generally implies that possibly risk-averse injurers bear risk. In some contexts, such as auto accidents, one supposes that injurers are not substantially less risk averse than victims.

[12] Some jurisdictions have implemented "no-fault" regimes (essentially, insurance that covers losses suffered in accidents) for automobile accidents. See Dewees, Duff and Trebilcock (1996). Also, there are intermediate schemes, like workers' compensation, that provide compensation and charge experience-rated premiums to injurers to instill incentives to reduce risk. See Moore and Viscusi (1990).

[13] See Danzon (1985, p. 187), Kakalik et al. (1983), and Shavell (1987a, p. 263).

[14] Farber and White (1991) provide evidence that many medical malpractice cases are dropped after discovery, when plaintiffs learn that the defendant probably was not negligent. Relatedly, Ordover (1978) analyzes a model in which victims are uncertain about injurers' negligence; the result is that some victims of negligence do not sue and others who are not victims of negligence do sue.

negligence rule due care will be at issue. In consequence, it is in theory ambiguous whether strict liability or the negligence rule will be administratively cheaper.

2.3.2. *Administrative costs and the social desirability of the liability system*

The existence of administrative costs and their significant magnitude raises rather sharply the question whether it is worthwhile for society to bear them to gain the benefits of the liability system – the incentives to reduce risk. Unfortunately, it is quite possible for suits to be attractive for private parties to bring even if the social benefits of the liability system are small and make it socially undesirable. For example, victims will have strong incentives to bring suit under a strict liability system however low the risk reduction effect of suit may be. This point about the private versus the social incentive to make use of the legal system will be emphasized in Section 5.1.2.

2.4. *Magnitude of liability: damages*

The magnitude of the payment a liable party must make is known as damages, because it is normally set equal to the harm the victim has sustained. In this section, we discuss various issues relating to damages.

2.4.1. *Basic theory*

As a general matter, damages should equal harm under strict liability for incentives to be optimal in the unilateral model of accidents. Clearly, for injurers to be led to choose optimal levels of care, their expected liability must equal expected harm $p(x)\,h$, meaning that damages d should equal h. Likewise, for their levels of activity to be optimal, the same must be true[15].

We should add that this point essentially carries over to the situation, not yet considered, where the magnitude of harm is stochastic. In this case, if damages d equal harm, then expected liability will equal expected harm, so incentives will be correct. However, if damages d do not equal actual harm but instead are set equal to $E_c(h)$, expected harm conditional on harm occurring, incentives will also be correct. (For elaboration, see Section 2.4.4.)

Under the negligence rule, analysis of the optimal magnitude of damages is somewhat different. Recall that if damages equal harm h, injurers will be induced to take care of x^* (assuming that due care $\hat{x}=x^*$). It is also the case that damages higher than h would induce injurers to take care of x^*: this will increase the incentive to be non-negligent, to choose x^*, but it will not lead injurers to take excessive care

[15] In the bilateral model, damages equal to harm would also be optimal under a rule of strict liability with a defense of contributory negligence if victims' activity level is not variable. If their activity level is variable, then optimal damages may well be less than harm, for this will induce victims to moderate their level of activity.

because they can escape liability merely by taking care of x^*. Moreover, it can be shown that damages somewhat below h will also induce due care because, by taking due care rather than slightly less care under the negligence rule, injurers do not just reduce liability slightly but avoid liability altogether[16]. Thus, optimal damages are not unique but range from a level somewhat below h to any greater level. When, however, one introduces the possibility of uncertainty in the negligence determination (see Section 2.1.1), the situation becomes more complicated. For example, we noted that error in the negligence determination might lead injurers to take excessive care to reduce the risk of being found negligent by mistake. If so, a level of damages exceeding h would only exacerbate this problem, and it might be beneficial for d to be lower than h[17].

To sum up, we can say that in simple cases damages should equal harm under strict liability and under the negligence rule, although there are complications, such as that concerning uncertainty in the negligence determination. In fact, the law generally does impose damages equal to harm, but subject to some exceptions (which we will note in Sections 2.4.3 and 2.4.5).

2.4.2. Nonpecuniary elements of loss

Accidents often involve nonpecuniary losses, such as pain and suffering. To provide injurers with proper incentives to reduce accidents, they should pay for all nonpecuniary harms that they cause. However, it may be better for the state to receive these payments than for victims to receive them. Victims would often not elect to insure against nonpecuniary losses because these losses would not create a need for money, that is, raise their marginal utility of wealth[18]. Parents usually would not insure against the death of a child, for example, as this frequently would not generate a need for money, however devastating the loss would be for the parents. Thus, as initially proposed by Spence (1977), liability for pecuniary losses accompanied by an appropriate fine for nonpecuniary losses may be socially desirable[19].

2.4.3. Punitive damages

When an injurer's behavior departs substantially from what is appropriate, damages in excess of harm, so-called punitive damages, may be imposed. If imposition of such

[16] This point depends upon the particular formulation of the negligence rule (whether a person who takes less than due care is responsible for all harm caused or only the increment to harm resulting from x falling below $\hat{x}$). See Kahan (1989).

[17] An additional issue is that erroneous findings of liability tend to remedy the problem of excessive levels of activity under the negligence rule, raising the possibility that setting damages above h would be desirable.

[18] For empirical evidence, see Viscusi and Evans (1990).

[19] Alternatively, victims might enter into contracts under which insurers would receive pain and suffering recoveries in exchange for a reduction in premiums on other coverage.

damages causes expected liability to exceed expected harm, injurers will be induced to take excessive precautions, at least under strict liability, and they will also reduce their levels of activity undesirably[20].

Damages exceeding liability are, however, desirable if injurers sometimes escape liability. This possibility arises because injurers may be hard to identify as the sources of harm (the origin of pollution may be difficult to trace) or because victims may not choose to bring suit (litigation costs may discourage legal action). If injurers who ought to be found liable for harm h are in fact only found liable and made to pay damages with probability q, then if damages are raised to $(1/q)h$, injurers' expected liability will be h. Thus, the more likely a party is to escape liability, the higher should be damages when the party is found liable. Accordingly, a firm that dumps toxic wastes at night, or an individual who tries to conceal a bad act, should have to pay punitive damages, but not an injurer who causes harm in a noticeable way. On these points and others, see for example Cooter (1989), Diamond (1997), and Polinsky and Shavell (1998a); and for empirical study, see Eisenberg et al. (1997), Karpoff and Lott (1999), and Polinsky (1997).

2.4.4. Accuracy of damages

Much expense is incurred in litigation about the magnitude of a victim's harm, which raises the question of what the social value of greater accuracy is and whether the private value of accuracy is different from the social value. As stressed in Kaplow and Shavell (1996b), the private value of accuracy about harm generally exceeds the social value. To explain, there is social value in establishing harm accurately primarily when injurers know, at the time that they choose their level of care, how much harm they might cause. For example, if an injurer anticipates that the atypically large harm he might cause will be accurately measured, he will exercise an appropriately high degree of care, as is socially desirable[21]. However, injurers often lack (and could not reasonably obtain) considerable information about the harm they might cause when they decide on their precautions. Drivers, for example, know relatively little about how much harm a potential victim would suffer in an accident (the seriousness of injuries, the magnitude of lost earnings). Thus, drivers' incentives to avoid accidents would be largely the same if, instead of using precise measurements of harm, courts employed rough averages (based, perhaps, upon abbreviated litigation over damages

[20] Under a perfectly operating negligence rule, punitive damages would not affect injurers' behavior, as explained in Section 2.4.1. But if there is uncertainty in the negligence determination, the problem of excessive precautions may be exacerbated by punitive damages; also, punitive damages may reduce injurers' activity levels (although this effect may be desirable).

[21] Relatedly, a prospective injurer's incentive to acquire information about the harm he may cause (whether it will be atypically large) will be greater when he knows that harm will be accurately determined.

or upon figures from a table)[22]. Nevertheless, victims and injurers have very strong incentives to spend to establish damages accurately in court. A victim will always be willing to spend up to a dollar to prove that harm is a dollar higher, and an injurer will always be willing to spend up to a dollar to prove that harm is a dollar lower.

2.4.5. Components of loss that are difficult to estimate

Some components of loss are hard to estimate, for example, the decline in profits caused by a fire at a store (as opposed to the cost of repairing the store) or certain nonpecuniary harms, and the law sometimes excludes such difficult-to-measure elements of loss from damages. This legal policy might be justified when the cost of ascertaining a component of loss outweighs the value of the improvement in incentives that its inclusion would accomplish. However, the cost of estimating a component of loss would be low if rough estimates were used (and the analysis of the last section suggests that this often would not much compromise incentives to reduce risk). Therefore, the policy of excluding components of loss that are hard to evaluate may be unwarranted.

2.5. Causation

2.5.1. Basic requirement of causation

A fundamental principle of liability law is that a party cannot be held liable unless he was the cause of losses. For example, if cancer occurs in an area where a firm has polluted, the firm will in principle be liable only for the cancer that it caused, not for cancer due to other carcinogens.

This principle is clearly necessary to achieve social efficiency under strict liability, because otherwise incentives would be distorted. Socially desirable production might be rendered unprofitable if the firm were held responsible for all cases of cancer.

Under the negligence rule, restricting liability to accidents caused by an actor may be less important than under strict liability: if negligent actors were held liable for harms they did not cause, they would only have greater reason to act non-negligently, but would not take excessive precautions if there were no uncertainty surrounding the negligence determination. In the presence of such uncertainty, however, relaxation of the causation requirement might adversely affect incentives. Further, under both liability rules, absence of the causation requirement might raise the volume of litigation

[22] A qualification to this point, emphasized by Spier (1994b), arises where the probability distribution of harm is affected by an injurer's degree of care. If this is so, then accuracy in assessing harm will influence an injurer's incentives to reduce risk even when, at the time he chooses his level of care, he does not have information about the harm that would occur in an accident.

and thus administrative costs. On the basic causation requirement and incentives, see originally Calabresi (1975) and Shavell (1980a)[23].

2.5.2. *Uncertainty over causation*

In many situations there is uncertainty about causation. For example, it may not be known which manufacturer out of many sold the product that resulted in injury, or whether harm was due to the defendant firm or to background factors (was cancer attributable to a firm's pollutant or to unknown environmental carcinogens?). The traditional approach of the law is to hold a defendant liable if and only if the probability that the defendant was the cause of losses exceeds 50%. This approach can lead either to inadequate or to excessive incentives to reduce risk. For example, a firm that supplies only 20% of the market demand will escape liability for any harm caused by its product (assuming that harm cannot be traced to particular firms). Consequently, the firm will have no liability-related incentive to take precautions. If, however, a firm's market share exceeds 50%, the firm will be liable for all harms due to the product that it and other firms sell, for it will always be correctly said to be more likely than not the cause of harm. Thus, the firm's liability burden will be socially excessive (under strict liability). These potential problems of inadequate and of excessive incentives may arise under any liability criterion based on a threshold probability of causation; they are not unique to a 50% threshold. Essentially this point has been made frequently, and it is formally developed in Shavell (1985b).

The legal system has recently adopted (in limited settings) the approach of imposing liability in proportion to the likelihood of causation. Under this approach, a firm supplying 20% of the market would be liable for 20% of harm in every case. Note, therefore, that the firm's liability bill would be the same under this regime as it would be if it paid for all the harm in the 20% of cases it truly caused – implying that its incentives would be socially appropriate. That the proportional liability principle engenders optimal incentives (without there being a need to establish causation in particular cases) is an advantage of the principle relative to the traditional threshold probability criterion. See Rosenberg (1984) and Shavell (1985b).

2.5.3. *Proximate causation*

Even if a party is a cause of losses, he may still escape liability under tort law because he was not the *proximate cause* of losses, where proximately-caused losses are, mainly, those that came about in an ordinary manner and that were not the product

[23] On the causation requirement under the negligence rule, see also Grady (1983) and Kahan (1989), who study restriction of liability to losses that are in excess of the possibly positive losses that the actor would have caused had he not been negligent.

of coincidence. Also, liability often is not found, on causal grounds, where accidents are freak events, such as where a dog imbibes nitroglycerin left at a mining site and then explodes, injuring nearby persons. Allowing parties to escape liability for such unusual accidents is sometimes thought not to undermine incentives, on the ground that no one could have foreseen such accidents. This argument, however, is subject to the criticism that courts may find it difficult to discriminate between accidents that can and cannot be foreseen. Moreover, the argument leads to the reductio ad absurdum that there should never be liability: any accident may be viewed as extraordinarily unlikely (of essentially zero probability) if it is described in sufficient detail.

The possibility that a party would not be said to be the proximate cause of losses on account of coincidence (as opposed to the freak character of losses) is illustrated by the following case: a speeding bus happened to be at just the "right" point on its route to be struck by a falling tree; the bus company escaped liability for the injuries to passengers even though they would not have occurred but for the excessive speed of the bus. Allowing parties to escape liability for such coincidental accidents might not affect precautions, however. One presumes that the probability of a bus being struck by a falling tree is independent of its speed, so that imposing liability would not affect the speed at which buses are driven. On proximate causation, see Calabresi (1975) and Shavell (1980a).

2.6. Judgment-proof problem

The possibility that injurers may not be able to pay in full for the harm they cause is known as the judgment-proof problem and is of substantial importance, for individuals and firms often pose risks significantly exceeding their assets (a person of modest means could cause a devastating fire; a small firm's product could cause many deaths). When injurers are unable to pay fully for the harm they may cause, their incentives to reduce risk will be inadequate and their incentives to engage in risky activities too great. See Shavell (1986).

It should be remarked as well that injurers who may not be able to pay for the entire harm they cause will tend not to purchase full liability insurance, or any at all. This is because purchase of full coverage will involve the purchase of coverage against a loss that a party would not fully bear in the absence of coverage: if a person with assets of $10 000 buys coverage against liability of $100 000, he is purchasing coverage against $90 000 of losses that he would not suffer if he did not have coverage. See Keeton and Kwerel (1984) and Shavell (1986).

Several types of policy response to the dilution of incentives caused by the judgment-proof problem are of interest. First, if there is another party who has some control over the behavior of the party whose assets are limited, then the former party can be held vicariously liable for the losses caused by the latter. Thus, holding a large contractor liable for the accidents caused by a small subcontractor or an employer for

accidents caused by its employees will induce the former to control the risks posed by the latter[24]. See Kornhauser (1982) and Sykes (1981).

Second, parties with assets less than a specified amount could in some contexts be prevented from engaging in an activity. However, such minimum asset requirements are a somewhat blunt instrument for alleviating the incentive problems under consideration.

A third response to inadequate incentives, one closely related to asset requirements, is regulation of liability insurance. See Shavell (2000). One form of insurance regulation would mandate purchase of (perhaps full) coverage[25]. This approach would be especially appealing when insurers can observe the precautions taken by injurers. An opposite form of insurance regulation would prohibit purchase of liability insurance. This could improve incentives to take care if insurers cannot observe injurers' precautions, because in that case insurance coverage would dilute incentives to take care when these incentives are inadequate to begin with.

Fourth, the use of Pigouvian taxes equal to expected harm may help to alleviate the judgment-proof problem. When harm will be caused with a low probability, the expected harm will be much less than actual harm; hence, parties with limited assets may be able to pay the appropriate tax on risk-creating behavior even though they could not pay for the harm itself.

A fifth way of correcting for dilution of incentives is for the state to regulate parties' behavior directly, such as with traffic laws or by insisting that food and drugs meet certain safety requirements. Regulation, however, may involve inefficiency because of regulators' limited knowledge of risk and of the cost and ability to reduce it. (We discuss regulation further in Section 3.6.2.)

A final way of mitigating dilution of incentives is resort to criminal liability. A party who would not take care if only his assets were at stake might be induced to do so for fear of imprisonment.

2.7. Product liability

We have not yet considered accidents where victims are customers of injurers (or, more generally, where victims are in some contractual relationship with injurers). In this case, the role of liability in providing incentives may be attenuated or even nonexistent. The reason, obviously, is that firms producing risky products may be unable to sell

[24] Imposing liability on corporations for behavior of their judgment-proof subsidiaries or requiring that liability of shareholders be unlimited (at least with respect to tort victims) might serve a similar function. See Hansmann and Kraakman (1991). Also, liability might be imposed on parties who supply services to potentially judgment-proof entities and are in a position to monitor them, such as accountants, lawyers, and lenders. See Feess (1999), Kraakman (1986), and Pitchford (1995).

[25] Many jurisdictions require liability insurance of those who drive, although the required amount is usually small.

them or may have to accept a reduction in price commensurate with the risk of loss attaching to the products.

If customer knowledge of product risk is perfect, then firms' incentives to reduce risk will be optimal even in the absence of liability. For example, if the expected losses caused by a product risk are \$100, the firm will have to accept a \$100 lower price than otherwise, so it will be willing to spend up to \$100 to eliminate the risk. Therefore, liability is not needed to generate incentives toward safety.

If, however, customer knowledge of risk is imperfect, liability is potentially useful in reducing risk[26]. In the absence of liability, firms that increase safety generally will be unable to obtain an increase in the price fully reflecting the reduction in risk[27]. (Indeed, in the extreme case where customers cannot observe anything about the true risk, firms would have no incentive to reduce it.) Therefore, the prospect of liability for product-caused harms will increase incentives to reduce risk. Also, imposing liability will result in prices that reflect the full costs of products, leading to more efficient purchasing decisions.

A question concerning liability is whether court-determined liability or customer-selected liability, namely, warranties, is likely to be better[28]. The answer depends on the nature of customers' information or lack thereof and on other factors. For example, suppose that customers cannot directly determine the risk associated with a product but realize that firms will minimize production costs plus expected accident losses if they have to bear those losses. Then, consumers may rationally elect to purchase a full warranty – essentially to adopt strict liability – because they know that the product with that warranty will really be cheaper than an apparently similar product sold without the warranty at a lower price. In this case, warranty selection leads to optimality.

Suppose instead that customers misperceive risk. Then their selection of warranties may be skewed, as emphasized by Spence (1977). For example, if customers believe the risk of a product failure causing a loss of \$10 000 to be 1% when it is really 5%, then a warranty would not be purchased: a seller of a full warranty would have to charge \$500 for it, but the perceived expected value of it would be only \$100. In this circumstance, it might be better for the courts to impose liability because that would create incentives to reduce risk. However, in many contexts, customers can significantly reduce product risks by exercising care in the use of products, and producer liability might dilute their incentives to do so (assuming that defenses such as contributory negligence are unsuccessful because of difficulties in observing customers' behavior). Also, the administrative costs of liability are high. Thus, whether imposition of liability will improve social welfare, given customers' ability to purchase warranties, involves a complicated weighing of considerations, and courts' ability to do this is not clear.

[26] See, for example, Goldberg (1974). Another potentially useful policy is supplying information about risk to customers, on which see Magat and Viscusi (1992) and Viscusi and Magat (1987).

[27] However, as emphasized by Schwartz and Wilde (1979), this point depends on the fraction of customers who are uninformed about the risk.

[28] See generally Priest (1981) and Rubin (1993).

2.8. *Liability versus other means of controlling accidents*

Liability is only one means of controlling harm-causing behavior; safety regulation and Pigouvian taxes are among the alternatives, as we indicate in Section 2.6[29]. For a general comparison of methods of controlling harm, see our discussion of regulating externalities in Section 3.6.2.

2.9. *Intentional torts*

To this point, we have examined liability for accidents, but we have not dealt explicitly with so-called intentional torts, such as assaulting someone or stealing his property (which also are crimes)[30]. See Landes and Posner (1981). An intentional tort may be defined as a harm that an injurer causes in which either of two things are true: the injurer acted in a manner that caused harm to occur with a very high probability, or the injurer obtained utility from the victim's suffering itself.

It would be possible to apply the foregoing analysis of accidents to intentional torts without modification. The conclusions reached did not depend on the magnitude of the probability of harm or on the source of benefits to injurers. However, both of these aspects of intentional harms suggest changes in assumptions that could alter our analysis and conclusions.

First, in situations where harm would be very likely to occur, bargaining between injurers and victims would often be possible[31]. If so, it may be desirable to forbid injurers from harming victims unless they obtain consent in advance, presumably in exchange for payment. Thus, thieves would be required to buy property they want, rather than simply take it and pay damages under a liability rule. On these issues, see Calabresi and Melamed (1972) and Kaplow and Shavell (1996a).

Second, where injurers derive utility directly from the fact that harm is suffered by victims, some analysts suggest that injurers' utility should not count in assessing social welfare. If so, deterrence becomes more valuable. Also, it may be optimal to deter some harmful acts even when the injurer's benefit exceeds the victim's loss, which calls for damages greater than harm, or for supplemental sanctions, notably imprisonment. (One suspects, however, that with most such intentional torts, injurers' benefits rarely exceed victims' losses.)

3. Property law

We begin our discussion by reviewing reasons why property rights should exist and by describing instances of their emergence. Then we consider the major questions

[29] For comparisons of liability and safety regulation, see Kolstad, Ulen and Johnson (1990), Schmitz (2000), Shavell (1984a), and Wittman (1977).

[30] On crime, see Section 6.4.

[31] We have implicitly ignored bargaining, except in our product liability discussion, because with most accidents – such as automobile accidents – bargaining between potential victims and potential injurers would be infeasible.

addressed by property law: the division and form of property rights, public property, the acquisition and transfer of property, and conflicts in the use of property (externalities). Last, we examine the subject of intellectual property. Many of the topics in this section have not been formally analyzed.

3.1. Justifications for property rights

A time-honored and fundamental question is why should there be any property rights in things[32]. That is, in what respects does the protection of property and the ability to transfer property promote social welfare? One justification for the protection of property is that it furnishes incentives to work, a common example being that people would not grow crops unless they could keep the product of their labor. Similarly, property rights provide incentives to maintain and improve durable things: to repair buildings, to fertilize and irrigate land, to conserve renewable resource stocks[33].

Another justification for property rights is that, were they absent, individuals would spend time and effort trying to take things from each other and protecting things in their possession, and they would often find themselves involved in conflict. Enforcement of property rights by the state, while involving its own costs, reduces these serious disadvantages that would be incurred in the absence of property rights. A related benefit of enforcing property rights is that it protects people against risk. In the absence of protection of property rights, individuals would face the possibility that their property would be taken from them (even though they might also enjoy the possibility that they would be able to take property from others).

In addition, it is important that a system of property rights allows for things to be transferred freely. Most obviously, if things can be traded, they will tend to be allocated to those who value them most[34]. Moreover, the ability to transfer things is indirectly necessary to our enjoyment of economies of mass production and specialization of labor, for when a large quantity of a good is produced by a single entity, the output ultimately will have to be distributed, which is to say, transferred, to many other individuals, and the entity will also often need to obtain inputs from other parties. In addition, transferability of property (particularly of land) allows it to be used effectively as collateral, thus enabling credit markets to function[35].

[32] A related question concerns how such rights should be protected. See Calabresi and Melamed (1972) and Kaplow and Shavell (1996a).

[33] Problems with conserving renewable resources that arise in the absence of property rights are often referred to as the tragedy of the commons. See Gordon (1954), Hardin (1968), Libecap (1998), and Ostrom (1998).

[34] We also note that protecting the security of property rights promotes the transfer of property: without protection of property rights, prospective buyers would not be inclined to buy things that might subsequently be stolen, and prospective sellers would be wary of making their ownership of valuable possessions known to others.

[35] Empirically oriented literature on the various benefits of property rights includes Alston, Libecap and Schneider (1996), Atwood (1990), Besley (1995, 1998), and Feder and Feeny (1991).

Early writing about property rights – by Bentham (1830), Blackstone (1765–1769), and Hobbes (1651), among others – stressed the justifications involving incentives to work and avoidance of strife. Today, the virtues of property rights seem to be taken for granted or are only casually asserted. Further, they are often conflated with the case for private property and the market system. This is a mistake, in that the various benefits from property rights that we mentioned could be enjoyed under a centrally planned economy. For example, incentives to work can be provided by paying workers on the basis of effort, even if a state enterprise owns what they produce. (Indeed, employees of profit-maximizing firms in private-enterprise economies are generally motivated by pay rather than by the literal ability to sell what they produce.) And the benefits of avoiding strife and theft might be enjoyed just as much under a centrally planned economy as under a market economy. The arguments for the social value of the market-enterprise system over central planning are different from those justifying the existence of property rights per se. (The arguments favoring market systems are based largely on the informational burdens that central planners face, problems of corruption, and the like.)

3.2. Emergence of property rights

We would expect property rights to emerge from a background of no rights or only poorly established rights when the various advantages of their existence substantially outweigh the costs of establishing and maintaining the rights[36]. Property rights will be likely to arise in these circumstances because, if many individuals recognize that they will probably be better off under a regime with property rights, pressures will be brought to bear to develop them.

Various examples of the emergence of property rights have been studied. Umbeck (1981) examines property rights during the California Gold Rush. When gold was discovered in California in 1848, property rights in land and minerals were largely undetermined and there were virtually no authorities to enforce the law. Almost immediately, however, arrangements were made to protect property rights in gold-bearing land and river beds. This encouraged individuals to pan for gold, to build sluices, and otherwise to invest to extract gold; it also curbed wasteful efforts to grab land and gold from one another.

An additional example of historical interest is the establishment by the Indians of the Labrador Peninsula of rights in land where none had existed. Demsetz (1967) connects this 17th-century event to the increased value of furs. He suggests that without property rights in land, overly intensive hunting of fur-bearing animals (especially beaver) would have taken place and the stock of animals would have been depleted.

[36] We note that property rights can be established and enforced by the state or informally, through social norms. On the latter, see Ellickson (1989, 1991) and Sethi and Somanathan (1996). Property rights also might be enforced by private organizations, such as the Sicilian mafia. See Gambetta (1993).

A more recent instance of the emergence of property rights concerns resources of the sea, as described in Biblowit (1991) and Eckert (1979). For most of history, there were no property rights in the ocean's fisheries because the supply of fish was inexhaustible for all practical purposes, but fish populations have come under strain with the use of modern fishing methods. To provide incentives to preserve fisheries, it has come to be accepted that countries have property rights in fish found in their coastal waters. Also, property rights have recently been established in the sea bed to foster exploration and extraction of oil and mineral resources. Another important example of the appearance of property rights concerns rights to the electromagnetic spectrum; assignment of these rights prevents garbling of signals and encourages investment in programming and transmission as well as trade of rights to high-value users. See DeVany et al. (1969) and McMillan (1994)[37].

3.3. Division and form of property rights

3.3.1. Division of property rights

From a conceptual viewpoint, what we speak of somewhat loosely as property rights can be divided into more basic rights: particular possessory rights, and rights to transfer these rights. A possessory right in a thing is the right to use it in a specified way at a named time and under a particular contingency. A right to transfer a possessory right is the right to give or sell a possessory right to another person. Thus, what we commonly conceive of as "ownership" of something (say, land) entails both a large swath of possessory rights (rights to build on land, plant on it, and so forth, under most contingencies, and into the infinite future) and associated rights to transfer them.

In fact, property rights in things are generally held in substantially agglomerated bundles, but there is also significant partitioning of rights contemporaneously, according to time and contingency, and according to whether the rights are possessory or are for transfer. For example, an owner of land may not hold complete possessory rights, in that others may possess an easement giving them the right of passage upon his land, or the right to take timber, or the right to extract oil if found (thus a contingent right). A rental agreement constitutes a division of property rights over time; wills provide for future and often contingent division of rights (depending on the survival of beneficiaries). Trust arrangements, such as those under which an adult manages property for a child, divide possessory rights and rights to transfer.

The division of possessory rights may be valuable when different parties derive different benefits from them, because gains can then be achieved if rights are allocated to those who obtain the most from them. There are, however, several disadvantages to the division of possessory rights or too fine a division of the rights. Individuals may

[37] Literature on the emergence of property rights is surveyed in a general discussion of property rights and economic activity in Libecap (1986).

wish to exercise the same rights at the same time (a person with a right of passage may wish to use a path that is currently blocked by the owner's use); externalities and related conflicts may arise (a person with a right of passage might trample crops). In addition, logistical problems may impede the division of rights (consider the problem of many individuals trying to share the use of a single automobile).

We also note that possessory rights and rights to transfer are ordinarily combined because this promotes efficiency: possessors will make appropriate investments if they are the ones who will benefit from subsequent sales, and possessors will ordinarily have superior knowledge about which opportunities for sale are most profitable. Sometimes, however, separation of possessory rights and rights to transfer may be beneficial. A child may own property but not have the right to sell it because an adult trustee can make decisions superior to those of the child; a renter of an apartment may not have the right to sublet it because he does not have sufficient reason to consider the character of another tenant (such as whether the tenant would be likely to disturb neighbors).

3.3.2. Consolidated form of property rights and the theory of the firm

Ownership of separate productive assets is often consolidated; namely, it is held by a single entity, the firm. The question of what constitutes the benefits of this form of ownership was initially posed by Coase (1937) and has subsequently been developed by, among others, Williamson (1975, 1985), Klein, Crawford and Alchian (1978), Grossman and Hart (1986), and Hart (1995)[38]. Here, we review the main factors that bear upon the relative advantages of separate versus consolidated holding of assets by firms[39].

First, consolidated ownership of assets reduces transaction costs because internal transfers of goods and services may be accomplished by command, eliminating the need for negotiation and bookkeeping expense[40]. Such reduction of transaction costs, however, often could be obtained as well by separate owners if they entered into long-term supply contracts, honored standing orders, and the like.

Second, consolidated ownership may lead to a dilution of incentives to work, in comparison to the situation where each individual owns the assets he uses in production. Firms can combat this incentive problem in two familiar ways: if they can observe individuals' efforts, they can penalize shirking; if not, they

[38] See also Alchian and Demsetz (1972), Hart (1989), Hart and Moore (1990), Holmström and Tirole (1989), and Jensen and Meckling (1976). For a discussion of different forms of consolidated ownership (including employee-owned firms, cooperatives, and nonprofits), see Hansmann (1996).

[39] The subject of consolidated versus separate ownership of productive assets could be viewed as falling under the heading of division of property rights (separate ownership being division of consolidated ownership), but we find distinguishing the two subjects helpful.

[40] This savings may involve some sacrifice. For example, information on the profitability of separate functions may be lost (unless there is internal transfer pricing, which may involve transaction costs similar to those of market exchange).

can tie compensation to measures of output[41]. Of course, both methods have costs. (Interestingly, the latter may re-introduce transactions costs, such as if transfer pricing is required to compute a manager's contribution to the firm's profits.)

Third, consolidated ownership enables a firm to avoid breakdowns in bargaining that would occur under separate ownership due to asymmetric information. For example, under separate ownership, the seller of a factor input might overestimate its value to the next-stage producer and demand too much for it, stymieing an efficient transfer. Under consolidated ownership, efficient transfers can be ordered[42]. Alternatively, however, separate owners could contract in advance for transfers to occur at a predetermined price.

Fourth, consolidated ownership may help to alleviate problems of inadequate investment in assets. An asset owner may not have a sufficient incentive to make a relationship-specific investment (upgrading a plant for producing a factor input) because he anticipates that his gains will be partially expropriated by the owner of a complementary asset at the time when he is to put his asset to use. But if both assets are owned by the same party, the problem of expropriation of the gains from the relationship-specific investment in the first asset will be mitigated, and investment in it should be more efficient. However, the other individual's incentive to invest in what otherwise would have been his asset may be dulled if the first party owns both assets; thus, consolidated ownership does not necessarily improve investment incentives overall. Additionally, it may sometimes be possible under separate ownership of assets to guarantee that investments in them be sufficient by making a contract to that effect; but this requires that investments be observable.

We close by noting that the distinction between consolidated ownership of assets by firms and separate ownership is blurred because, as we have mentioned, under separate ownership together with contractual arrangements, it is often possible to replicate the advantages of firms. Indeed, separate ownership combined with sufficiently encompassing contracts may be indistinguishable from the consolidation of ownership of assets by firms. Conversely, firms themselves can be understood to consist of a set of contracts (a corporation is a particular contract among its shareholders).

[41] Ellickson (1993), among others, suggests that most communal farming efforts have failed because individual rewards were not linked to effort or output, which led to widespread shirking. It may be observed as well that modern firms succeed despite their often large size through monitoring of workers' behavior (which may be more feasible with the use of mechanized technology) and use of performance pay.

[42] The manager might know that the transfer is efficient without knowing the precise cost and/or value of the factor input, for the cost distribution may be below the value distribution. When distributions are, instead, substantially overlapping, a manager will not know whether a transfer is efficient, and in this case bargaining between separate parties may well promote efficiency.

3.4. Public property

Before continuing with our analysis of property rights, we consider briefly an important class of property: that owned by the public. We review the justifications for public property and then two methods of acquisition of such property: by purchase and by unilateral public taking.

3.4.1. Justifications for public property

The main justifications for public property concern problems with private supply. The government builds and maintains roads, for example, because private supply often would not be forthcoming due to difficulties that would be faced in collecting for road use. And even if roads were privately supplied, suppliers would charge tolls, raising problems of monopoly pricing and wasteful expenditures on toll-collecting.

Problems with private supply, however, do not constitute an argument for public ownership of goods, only for public financing of them or for public regulation of private suppliers. A road could be constructed, maintained, and owned by a private party paid by the state. And when private ownership might involve problems of monopoly pricing, government regulation is an alternative to direct ownership. These observations underlie the growing attention to privatization of public property and of government activities. The comparative virtues of public versus private ownership depend on the relative abilities of the government and of the private sector to operate efficiently and maintain quality[43].

3.4.2. Acquisition of public property: purchase versus compensated takings

The state may acquire property through purchase or through exercise of the state's power of eminent domain, which is to say, by taking the property. In the latter case, the law typically provides that the state must compensate property owners for the value of what has been taken from them, and it will be assumed that this is the case until the next section.

The difference between purchases and compensated takings is that the amount owners receive is determined by negotiation in the former case but unilaterally by the state in the latter case. Because of possible errors in governmental determinations as well as concerns about the behavior of government officials, purchase would ordinarily be superior to compensated takings. An exception, however, arises where the state needs to assemble many contiguous parcels, such as for a road. Here, acquisition by purchases might be delayed or prevented by hold-out problems, making the power to take socially advantageous.

The actual pattern of governmental acquisition of property largely reflects these simple observations. Most state acquisition of real estate, and virtually all acquisitions

[43] See for example Hart et al. (1997), Shleifer (1998), and Viscusi et al. (1995, pp. 468–470).

of moveable property, is through purchase. Governmental takings are restricted mainly to situations where there is a need for roads, dams, and parks, and to establish certain private rights-of-way, such as for railroads or utility lines[44].

3.4.3. Compensation for takings

Assuming that there is a reason for the state to take property, consider the effects and desirability of a requirement that the state pay compensation to property holders. As emphasized by Blume et al. (1984), payment of compensation to property owners creates a potential moral hazard: it leads them to invest excessively in property. For example, a person may build a home on land that might be taken by the state for use for a road because he will be compensated for the home if the land is taken. However, building the home might not be socially justified, given the probability of use of the land for a road, which would require destruction of the home.

A second effect of compensation for takings is that risk-averse property owners will bear less risk[45]. But were takings not compensated, insurance against takings would be likely to emerge. Moreover, private insurance would naturally alleviate the problem of excessive investment in property[46].

Third, payment of compensation also may alter the incentives of public authorities to take property by reducing possible problems of overzealousness and abuse of authority. However, requiring compensation may also exacerbate potential problems of too little public activity (public authorities do not directly receive the benefits of takings). Therefore, it is not clear whether a compensation requirement improves the incentives of public authorities. For further discussion of these various issues about compensation for takings, see Kaplow (1986a, 1992a)[47].

3.5. Acquisition and transfer of property

We return now to the subject of private property and consider a number of topics relating to its acquisition and its transfer.

3.5.1. Acquisition of unowned property

Wild animals and fish, long-lost treasure, certain mineral and oil deposits, and, historically, unclaimed land, constitute primary examples of unowned property that

[44] See Bouckaert and De Geest (1995) on the related topic of private takings.

[45] It should be noted, however, that many property owners – namely, firms with diversified ownership – are not very risk averse.

[46] Notably, insurance premiums would be based on the value of property, so further investments would raise premiums.

[47] On the topic of compensation for loss in value of property due to regulation (as opposed to the physical taking of property), see Fischel (1995) and Miceli and Segerson (1996).

individuals may acquire. The law has to determine under what conditions a person will become a legal owner of such previously unowned property, and a general legal rule is that anyone who finds, or takes into his possession, unowned property becomes its owner.

Under this finders–keepers rule, incentives to invest in capture (such as to hunt for animals or explore for oil) are optimal if only one person is making the effort. However, if, as is typical, many individuals seek unowned property, they will invest a socially excessive amount of resources in search: one person's investment or effort usually will not simply increase the total probability of success, but rather will come, at least partly, at the expense of other persons' likelihood of finding unowned property[48].

Various aspects of the law governing the acquisition of property may be regarded as ameliorating this problem of excessive search effort under the finders–keepers rule[49]. Notable examples are that regulations may limit the quantities that can be taken of fish and wild animals, the right to search for oil and minerals on the ocean floor may be auctioned, and oil extraction may be "unitized" (assigned to one party)[50].

3.5.2. Loss and recovery of property

When property is lost by its owner and is found by another person, the question arises whether the original owner should retain property rights or the finders–keepers rule should apply. The general stance of the law is that original owners maintain their property rights in lost things (unless they abandon them). This beneficially discourages original owners from socially excessive investment in preventing losses: a farmer might otherwise invest in an expensive fence to prevent his cattle from straying, which might be inefficient because often his private loss would not constitute a social loss (someone would be likely to find the strays). Moreover, original owners usually can either search themselves or efficiently organize recovery efforts by others (including by offering rewards). If, however, original owners cannot do this, the finders–keepers rule does have the advantage of inducing recovery effort, even though the rule tends to encourage races to find the effectively unowned property. In any event, if original owners retain property rights, finders may simply hide what they find, which reduces the value of what is found without producing the aforementioned benefits to original owners.

3.5.3. Acquisition of stolen property and problems of establishing valid title

A basic difficulty associated with sale of property that a legal system must solve is establishing validity of ownership, or "title". How does the buyer know whether the

[48] This problem is similar to the tragedy of the commons. See Gordon (1954) and Hardin (1968).

[49] For a survey of relevant literature, see Lueck (1998).

[50] Other laws limit indirectly how much property can be taken by individuals by giving them title only if they make productive use of the property that they find. This was true of homestead laws that gave land to individuals who worked it and of water rights regimes that gave priority to the extent that water supplies were regularly used. Such rules, however, create excessive incentives to exploit property.

seller has good title, and how does the buyer obtain good title? If these questions are not readily answered, sales transactions are impeded, and theft may be encouraged.

One route that legal systems may take involves the use of registration systems: lists of items and their owners. Important examples are registries of land, ships, motor vehicles, and many financial instruments. Presuming that an item is recorded in a registry, it will be easy for a buyer to check whether the seller holds good title to it, and the buyer will obtain title by having his name recorded in the registry as the new owner. Also, a thief obviously cannot claim that something he has stolen is his if someone else's name is listed as the owner in the registry. Registries are usually publicly established, and listing in registries often is mandatory (or it may be encouraged by making registration a condition to asserting a valid legal claim). Partial explanations for the public role in registries are the coordination problem that may be involved in creating them and the problem of insufficient private incentives to register property to provide a general deterrent against theft. (An individual contemplating registration will not take into account that, as the proportion of registered property rises, thieves anticipate that it will be more difficult to sell stolen property and thus are discouraged from theft.)

For most goods, however, registries do not exist because of the expense of establishing and maintaining them relative to the value of the goods and of the deterrence of theft. Two legal rules for determination of title are available (and both, to some extent, are employed) in the absence of registries. Under the original ownership rule, the buyer does not obtain good title if the seller did not have it; the original owner can always claim title to the item if he can establish his prior ownership. Under the bona fide purchaser rule, a buyer acquires good title as long as he had reason to think that the sale was bona fide (that the seller had good title) – even if the item sold was in fact previously stolen or otherwise wrongfully obtained. These rules have different effects on incentives for theft. Notably, under the bona fide purchaser rule, theft is made attractive because thieves will often be able to sell their property to buyers (who will be motivated to "believe" that the sale is bona fide); the buyers can use the now validly held property or resell it. Another social cost of the bona fide purchaser rule is that original owners will spend more to protect their property against theft because theft will be more frequent and, when it occurs, owners will be less likely to recover their property. (These costs of protection, note, are analogous to those arising under the rule allowing finders of lost property to keep it.) Finally, under the bona fide purchaser rule, buyers will not have an incentive to expend effort determining whether there exists a third-party original owner. This is an advantage in the direct sense that it reduces transaction costs, but it also compromises deterrence of theft.

3.5.4. Involuntary transfer of property: adverse possession

The legal doctrine of *adverse possession* effectively allows involuntary transfer of land (and some other types of property): a person who is not the owner of land is deemed to become its legal owner if he takes possession of it and uses it openly and

continuously for at least a prescribed period, such as ten years. Some have suggested that a rationale for the rule is that it permits the transfer of land from those who would leave it idle to those who will use it productively. But this overlooks the possibility that there may be good reasons for allowing land to remain idle (perhaps it will be built upon later, and thus an investment in it now would be a waste). Furthermore, a prospective adverse possessor could always bargain with the owner to rent or buy the land. Additionally, the rule suffers from the disadvantage that it induces landowners to expend resources policing incursions onto their land and it encourages others to attempt adverse possession. (Observe that these latter arguments are similar to those in the preceding sections that favored rules protecting original owners.)

A historical justification for the rule is that, before reliable land registries existed, it allowed a landowner to establish good title to a buyer relatively easily: the seller need only show that he was on the land for the prescribed period. Another advantage of the rule is that it reduces disputes that would arise where structures turn out to encroach on neighboring parcels[51].

3.5.5. Constraints on sale of property

Legal restrictions are often imposed on the sale of goods and services. One standard justification for such policies is externalities. For example, the sale of handguns may be made illegal because of the externality that their ownership creates, namely, crime, and a tax may be imposed on the sale of a fuel because its use pollutes the air. See Section 3.6. The other standard justification for legal restrictions on sale is lack of consumer information. For instance, a drug may not be sold without a prescription because of fear that buyers would not use it appropriately. Here, though, one must compare the alternative of the government supplying relevant information to consumers (say that the drug has dangerous side effects, or that it should only be taken with the advice of a medical expert)[52].

3.5.6. Gifts

The making of gifts, including bequests, is the major way in which property changes hands other than by sale. Gifts are, as one would expect, rather freely permitted because, like sales, they typically make both involved parties better off[53]. It should be observed that, in the absence of a state subsidy, the level of giving may well fall short of the socially optimal level because a donor's private incentive to make a gift does not take into full account the donee's benefit. See Kaplow (1995b)[54]. In addition, some

[51] On adverse possession, see Netter (1998).

[52] For further discussion, see Section 4.1.9 on legal overriding of contracts.

[53] There are some limits on disinheriting one's immediate family and other rules that prevent individuals from controlling the use of their gifts long into the future (the rationale for which is not entirely clear).

[54] See also Friedman (1988) on the gift externality.

gifts, particularly to charities, may support public goods or accomplish redistribution, which may provide a further ground for subsidy[55]. In fact, the law does favor certain types of giving by conferring tax advantages on donees (and, in the case of charities, on donors). On the other hand, heavy gift and estate taxes are levied on large donative transfers to individuals.

Another issue concerning gifts is that a person may want to make a transfer in the future, in which case issues concerning contracts to give gifts arise. This subject will be discussed in Section 4.3.2 on donative contracts.

3.6. Conflicts in the use of property: externalities

3.6.1. Socially optimal resolution of externalities

When individuals use property, they may cause externalities, namely, harm or benefit to others. As a general matter, it is socially desirable for individuals to do more than is in their self-interest to reduce detrimental externalities and to act so as to increase beneficial externalities.

It should be noted, as emphasized by Coase (1960), that the socially optimal resolution of harmful externalities often involves the behavior of victims as well as that of injurers (and similarly with regard to generators of positive externalities and beneficiaries). Where victims can do things to reduce the amount of harm (install air filters to avoid pollution) more cheaply than injurers, it is optimal for victims to do so. Moreover, victims can sometimes alter their locations to reduce their exposure to harm. When the latter possibility is not incorporated into the analysis of externalities (suppose that victims are assumed to continue to live adjacent to a hazardous waste site), what is referred to as the optimal resolution of externalities may only be conditionally optimal.

3.6.2. Resolution of externalities through state intervention

We now consider various means of government intervention, along the lines of Shavell (1984a,c, 1993a)[56]. For convenience, we confine our attention to the case of harmful externalities, and we assume (until the next section) that parties affected by externalities cannot bargain with the generators of externalities.

Under direct regulation, the state restricts permissible behavior. It might impose a quantity constraint (a fisherman may be required to limit his catch to alleviate depletion of the fishery) or other behavioral constraints (a factory may be required to use a smoke scrubber). Closely related to state regulation is privately-initiated regulation through use of the legal *injunction,* whereby a potential victim can enlist the power of the state to force a potential injurer to take steps to prevent harm or to cease his activity.

[55] See, for example, Atkinson (1976) on redistribution and charitable contributions.

[56] See also Bovenberg and Goulder (2002, Chapter 23 of this Handbook) on environmental taxation.

Society can also make use of financial incentives to induce injurers to reduce harmful externalities. Under the Pigouvian tax, a party pays the state an amount equal to the expected harm he causes, for example, the expected harm due to a discharge of a pollutant into a lake. An additional type of financial incentive is a subsidy, an amount paid by the state to a party equal to the reduction in expected harm from some benchmark level that he accomplishes.

There is also liability – a privately-initiated means of providing financial incentives – as we discuss in Section 2. Under strict liability, a party who causes harm has to pay the victim for his losses. (Such liability differs from the corrective tax because payment is to the victim rather than to the state, and also because injurers pay for actual harm rather than for expected harm.) Under the negligence rule, an injurer must pay the victim only if the injurer failed to take a cost-effective precaution.

In fact, liability and regulation are the preeminent tools that society uses to control externalities; the use of corrective taxes and subsidies is unusual. Since Pigou (1932), who first emphasized the problem of externalities, economists have focused on corrective taxes and regulation, essentially ignoring liability. We will now sketch some factors bearing on the relative desirability of these methods of controlling externalities. The review of factors will show that any of the methods (or a combination) could be the best, depending on the context.

One factor of relevance is the quality of the state's information. If the state has complete information about acts, that is, it knows the injurer's benefit or cost of precautions along with the victim's harm, then all of the approaches allow achievement of optimality. But if the state's information is imperfect, it will not be able to calculate which actions (such as installing a smoke scrubber) are desirable and thus sometimes will err. However, if the state knows the expected harm, it can induce injurers to act optimally under the corrective tax or a rule of strict liability, because the injurer, who is presumed to know the cost of a precaution, will then appropriately balance the cost against the reduction in expected harm that would be brought about[57].

We emphasize that this basic informational argument favoring Pigouvian taxes or strict liability over regulation or the negligence rule extends to the case where the state is uncertain about the magnitude of harm. The reason, essentially, is that under the former rules, the state only needs to estimate expected harm (as the injurers themselves implicitly supply complete information about the costs of precaution when making their decisions). By contrast, under regulation and the negligence rule the state must estimate *both* expected harm and precaution costs. Because the state's effectively available information is strictly better under the corrective tax or strict liability, it can achieve a superior outcome. (This point holds notwithstanding Weitzman's argument

[57] This advantage, as it applies to the comparison between strict liability and the injunction, is suggested by Calabresi and Melamed (1972) and is further explored in Kaplow and Shavell (1996a) and Polinsky (1980b).

suggesting that quantity regulation may be superior to corrective taxation[58].) An implication is that the use of pollution taxes is superior to the use of tradeable pollution permits because, under the latter, the government sets the total quantity of pollution using its own estimate of abatement costs rather than implicitly relying on firms' information[59].

A second factor is the information available to victims. For many externalities, victims have better information than the state about who is causing harm or about its extent – because they actually suffer the harm – so they are the most appropriate enforcement agents, suggesting the desirability of the liability tool or the injunction. In other instances, however, victims may be unaware of the harm or its cause, making the state a better enforcer. State enforcement, such as by regulation or by corrective taxes based upon statistical evidence of expected harm, avoids the need to identify, say, which pollutants ultimately harmed which victims.

A third factor concerns the level of activity of an injurer (how much a firm produces, how many miles a person drives), as opposed to the precautions an injurer takes given the level of activity (whether a firm uses a smoke scrubber while producing, whether a person exercises care when driving). Regulation and the negligence rule are most often concerned with precautions taken but not with the level of activity: a factory may be required by regulation to install smoke scrubbers but not to reduce its output. Thus injurers may not have incentives to moderate their level of activity although that would be desirable (their activity may result in harm despite the exercise of optimal precautions – even with smoke scrubbers, some pollution will result). By contrast, under the corrective tax and strict liability, injurers pay for harm done, so that they will optimally moderate their level of activity (as well as efficiently choose their level of precautions).

A fourth pertinent factor, noted above, is the ameliorative behavior of victims. Under regulation, corrective taxation, and other approaches that do not compensate victims for their harm, victims have a natural incentive to take optimal precautions (or to relocate) because they bear their residual losses; they will want to take any precaution (install air filters to reduce pollution) whose cost is less than the reduction in harm it accomplishes. Under a strict liability rule, however, a victim might not have such an incentive because he would be compensated for his losses. But under a negligence rule, victims are not compensated if injurers have behaved properly, and, under strict

[58] Weitzman's (1974) conclusion that regulation could be superior to taxation rests on his assumption that the state must, in advance, set a corrective tax rate that is independent of the quantity of pollution. Yet, when the marginal harm depends on the quantity of pollution, the optimal tax rate depends on the quantity of pollution. See Roberts and Spence (1976). Kaplow and Shavell (2002b) emphasize that taxes that depend on quantity are usually feasible to implement and are superior to quantity regulation.

[59] To be sure, tradeable permit regimes are themselves superior to quantity constraints imposed at the level of individual firms because trading allows a given total pollution target to be reached at minimum cost.

liability, compensation might be given only to victims who took optimal precautions (if this can be determined)[60].

Still another factor is administrative costs, the costs borne by the state in applying a legal rule and the legal and related costs borne by the affected parties (aside from direct costs, such as the costs of precautions). Liability rules possess a general administrative cost advantage over regulation in that under liability rules, administrative costs are incurred only if harm is done. This advantage may be significant when the likelihood of harm is small. Nevertheless, administrative costs will sometimes be lower under other approaches. For example, compliance with a regulation may readily be detected in some circumstances (determining whether factory smokestacks are sufficiently high would be easy) and also may be accomplished through random monitoring, saving enforcement resources. Also, imposing corrective taxes might be inexpensive. Notably, suppose that they are levied at the time of the purchase of a product. In contrast, liability rules might be expensive to employ. For example, demonstrating the source of a particular harm and its extent may be difficult. Also, when industrial pollution affects millions of individuals on an ongoing basis, the cost of a continuous flow of individual suits (or even class actions) that measure damages victim-by-victim is likely to be in excess of the cost of alternatives.

Last, the ability of injurers to pay for harm is of relevance. For liability rules to induce potential injurers to behave appropriately, injurers must have assets sufficient to make the required payments; otherwise they will have inadequate incentives to reduce harm, as discussed in Section 2.6. Where inability to pay is a problem, bonding requirements may be helpful, and regulation may become more appealing (although it may need to be enforced through the threat of nonmonetary, criminal sanctions). In addition, corrective taxes have an advantage over liability rules when harm is probabilistic because, under the corrective tax, an injurer would pay only the expected harm (with certainty) rather than the actual harm (if there is a 1% chance of causing $1 000 000 of harm, the payment would be only $10 000). Many firms that would be able to pay the tax and thus have correct incentives would not be adequately deterred under a liability rule, on account of their inability to pay for harm when it actually occurs.

3.6.3. Resolution of externalities through bargaining by affected parties

Parties affected by unregulated externalities will sometimes have the opportunity to make mutually beneficial agreements with those who generate the externalities. In the classic example, if a factory's pollution causes harm of $1000 that can be prevented by installing a smoke scrubber that costs $100, then, in the absence of any legal obligation on the factory, one might expect a potential victim of pollution to pay the factory to install the scrubber. An agreement for any amount between $100 and $1000

[60] For further discussion of this aspect of liability rules, see Section 2.1.

would be mutually beneficial. Let us first consider this possibility and then evaluate its significance[61].

If it is posited that there are no obstacles to reaching a mutually beneficial agreement concerning externalities, then that will occur. This tautology is one version of the Coase Theorem; Coase (1960) stressed the point that externality problems could be remedied through private bargains. A closely related version of the Coase Theorem asserts that the outcome regarding the externality – whether a smoke scrubber is installed or instead pollution is generated – does not depend on the legal rule that applies. For example, if the scrubber costs $100 and there is no law that controls pollution, a bargain as we have described it will come about and the scrubber will be installed; and likewise if there is a law that leads to installation of the scrubber, the same will happen[62]. The outcome, however, might be affected by the legal rule because of the level of wealth of parties. Most obviously, the potential victims might not have assets sufficient to pay for the scrubber, in which case the scrubber would not be installed unless a legal rule leads to this; moreover, legal rules may affect the distribution of wealth and thus the demand for goods, including that of being free from pollution[63].

There are, however, many obstacles to bargaining. Bargaining may fail to occur when victims are numerous and face collective action problems in coming together. This is often the situation with respect to victims of industrial pollution. Similarly, in important contexts, bargaining will be impractical because victims will not know in advance who will injure them; this is the case for automobile accidents and most other accidents between strangers. Another reason that bargaining may not occur is that victims might not know that they are exposed to a risk (such as from an invisible carcinogen). Also, of course, the cost of bargaining between just one potential victim and one potential injurer who know of each other can discourage them from engaging in the process. If these reasons do not apply and victims and injurers do engage in bargaining, asymmetry of information may lead to bargaining impasses; for example, where a victim thinks that a smoke scrubber would cost a factory only $50 and it really costs $100, he may offer too little to the factory to reach an agreement. In all, these problems that reduce the likelihood of bargaining occurring, and also its success if it does take place, make the importance of legal rules to remedy externalities substantial.

[61] For experimental studies on bargaining and entitlements, see, for example, Hoffman and Spitzer (1982) and Croson and Johnston (2000).

[62] Similarly, if there is a law permitting victims to enjoin factories from polluting but pollution does less harm than it costs to prevent, the factory would pay the victim to forgo the injunction, resulting in the same outcome – pollution – as would occur with no regulation of pollution.

[63] The outcome following from a legal rule might also be affected by an "endowment effect", wherein individuals' valuations depend on whether or not they originally enjoy legal protection. See Kahneman, Knetsch and Thaler (1990).

3.7. Property rights in information

Legal systems accord property rights in information, including inventions, books, movies, television programs, musical compositions, computer software, chip design, created organisms, and trademarks. The generation and use of such information, and therefore the law governing it, is growing increasingly important in modern economies. We divide our review of this subject into three parts: First, we discuss certain information like an invention that can be used repeatedly to produce something; here we discuss patent, copyright, and trade secret law. Second, we examine diverse other types of information and its legal protection. Third, we consider labels of various types and their protection under trademark law.

3.7.1. Inventions, compositions, and other intellectual works of repeat value

The classic forms of intellectual works that receive legal property rights protection are inventions and literary, musical, or other artistic compositions. The well-known description of socially optimal creation and use of such intellectual works is as follows. First, it is socially optimal for an intellectual work, if created, to be used by all who place a value on it exceeding the marginal cost of producing or disseminating the good (or service) embodying it; thus a new mechanical device should be used by all who place a value on it exceeding the cost of its manufacture, and a book by all who value it more highly than its printing cost. Second, an intellectual work should be created if the cost of doing so is less than its total value to the public, net of production cost.

Given this description of social optimality, the advantages and disadvantages of property rights in intellectual works are apparent. In the absence of property rights, a creator of an intellectual work will obtain profits from it only for a limited period – until competitors are able to copy the creator's work. Thus, the generation of intellectual works is likely to be suboptimal. But if there exist property rights, whereby a creator of an intellectual work obtains a monopoly in goods embodying the work, incentives to produce the works will be enhanced (although they will still be less than ideal because innovators do not capture all of the surplus that their works create)[64]. The major drawback to intellectual property rights, however, is that monopoly pricing leads to socially inadequate production and dissemination of intellectual works[65]. This problem can be severe where the monopoly price is much higher than the cost of production. A good example is computer software, which may be sold for hundreds of dollars a copy even though its cost of dissemination is essentially zero. Another problem (with

[64] Kitch (1977) emphasizes a somewhat different view, under which patent rights are often granted at an early stage of invention, and the rights allow their holders to develop the inventions into commercially viable products.

[65] Relatedly, subsequent innovators whose inventions depend on prior patented works will need to obtain licenses from existing patent-holders, and hold-up problems may arise. See Chang (1995), Green and Scotchmer (1995), and Heller and Eisenberg (1998).

patent rights in particular) is the race to be the first to develop intellectual works. Given that the rights are awarded to whoever is first, a socially wasteful degree of effort may be devoted to winning the race, for the private award of the entire monopoly profits may easily outweigh the social value of creating a work before a competitor does[66].

Patent law and copyright law are the most familiar forms of legal intellectual property right protection[67]. The extent of protection afforded by each body of law is partial in various dimensions, however, so that they might be considered to represent a compromise between providing incentives to generate intellectual works and mitigating the monopoly problem. Patents and copyrights are limited in time (usually 20 years for patents, and the author's lifetime plus 50 years for copyrights) and also in scope[68]. As an example of the latter, the copyright doctrine of fair use often allows a person to copy short portions of a copyrighted work. This probably does not deny the copyright holder significant revenues (a person would be unlikely to purchase a book just to read a few pages), and the transaction costs of the copier having to secure permission would be a waste and might discourage his use.

A distinct form of legal protection is trade secret law, comprising various doctrines of contract and tort law that serve to protect not only processes, formulas, and the like that might be protected by patent or copyright law, but also other commercially valuable information such as customer lists. An example of trade secret law is the enforcement of employment contracts stipulating that employees not use employer trade secrets for their own purposes. A party can obtain trade secret protection without having to incur the expenses and satisfy the legal tests necessary for patent or copyright protection. Also, trade secret protection is not limited in duration (Coca-Cola's formula has been protected for over a century). However, trade secret protection is in some respects weaker than patent protection; notably, it does not protect against reverse engineering or independent discovery. On the economics of trade secret law, see Friedman, Landes and Posner (1991).

An interesting and basic alternative to property rights in information is for the state to offer rewards to creators of information and for information that is developed to be made available to all who want it[69]. Thus, under the reward system, an author of a book would receive a reward from the state for the writing of the book – possibly based on sales of the book – but anyone who wanted to print it and sell it could do so. Like the property rights system, the reward system encourages creation of information because the creator gains from producing intellectual works. But unlike the property rights system, the reward system results in the optimal dissemination of information

[66] The economic literature on intellectual property, focusing on patents, is discussed in Scherer and Ross (1990) and Tirole (1988); see also the historical review in Machlup (1958) and Reinganum's (1989) survey on the timing of innovation.

[67] See Besen and Raskind (1991), Gordon and Bone (2000), Landes and Posner (1989), and Menell (2000).

[68] See Gilbert and Shapiro (1990), Kaplow (1984), Klemperer (1990), and Scotchmer (1996, 1999).

[69] See Kremer (1998), Shavell and van Ypersele (2001), and Wright (1983).

because the intellectual works are placed in the public domain; anyone may use them for free. Hence, the reward system may seem to be superior to the property rights system. A major problem with the reward system, however, is that the state needs information about the value of innovations to determine rewards. We note that, to some degree, society does use a system akin to the reward system in that it gives grants and subsidies for basic research and for other intellectual works. But society does this largely when these intellectual works do not have direct commercial value.

3.7.2. Other types of information

There are many types of information different from what we have discussed above. One type of information is that which can be used only a single time, for example, where oil is located under a particular parcel of land. With regard to this type of information, there is sometimes no need for property rights protection. If the party who possesses the information can use it himself (to extract the oil), then once he does so, the issue of others learning it becomes moot – there will be no further value to the information. To the degree, though, that the party is unable to use the information directly (perhaps he cannot conveniently purchase drilling rights), his having property rights in the information might be valuable and beneficially induce the acquisition of information[70]. Moreover, we observe that giving property rights in the information will not undesirably reduce the use of information when the optimal use of it is only once. In fact, the legal system usually does furnish property rights protection in such information as where oil is located through trade secret law and allied doctrines of tort and contract law[71].

Another type of information is that relevant to future market prices. Here, the private and the social value of gaining such information can diverge, as emphasized by Hirshleifer (1971). For example, a person who first learns that a pest has destroyed much of the cocoa crop and that cocoa prices are therefore going to rise can profit by buying cocoa futures. The social value of his information inheres principally in any beneficial changes in non-financial behavior that it brings about. For example, an increase in cocoa futures prices might lead candy producers to reduce wastage of cocoa or to switch from chocolate production to production of another kind of candy. But the profit that a person with advance information about future cocoa prices makes can easily exceed its social value (suppose he obtains his information only an hour before it would otherwise become available, so that it has no social value) or fall short of its social value (suppose that he obtains information early on, but that his profits are low because he has limited funds to invest in futures). Hence, it is not evident whether it is socially desirable to encourage acquisition of such information about

[70] In addition, firms may need to be able to prevent employees from diverting a firm's benefit to themselves.

[71] See also our discussion of disclosure in Section 4.1.2 on contract law.

price movements by giving individuals property rights in the information. The law does not generally discourage such information acquisition (but an exception is regulation of trading based on insider information), and the law often encourages acquisition through trade secret protection[72].

Last, consider information of a personal nature about individuals. The cost of acquiring this information is the effort to snoop, although the information is sometimes adventitiously acquired, so costless. The social value of the information involves various complexities. The release of information of a personal nature to the outside world generally causes disutility to those persons exposed and utility for others, the net effect of which is ambiguous. Further, a person's behavior may be affected by the prospect of someone else obtaining information about him: he may be deterred from socially undesirable behavior (such as commission of crimes) or from desirable but embarrassing-if-publicly-revealed behavior, and he may make costly efforts to conceal his behavior. Thus, there are reasons why the acquisition and revelation of personal information are socially undesirable, and reasons as well why they might be socially beneficial. The law penalizes blackmail and in this way attempts to discourage profit from acquisition of personal information[73]. But otherwise the law does not generally retard the acquisition of personal information, and it also extends limited property rights in such information; notably, an individual who wants to sell to a publisher personal information he has obtained usually can do so.

As this brief discussion has illustrated, the factors bearing on the desirability of protecting property rights in information vary significantly according to the type of information and call for analysis quite different from that concerning information of repeat value that we considered above.

3.7.3. Information valuable as labels

Many goods and services are identified by labels. The use of labels has substantial social value because the quality of goods and services may be hard for consumers to determine directly. Labels enable consumers to make purchase decisions on the basis of product quality without going to the expense of independently determining their quality (if this is even possible). A person who wants to stay at a high-quality hotel in another city can choose such a hotel merely by its label, such as "Ritz Hotel"; the consumer need not directly investigate the hotel. In addition, sellers who label their output will have an incentive to produce goods and services of quality because consumers will recognize quality through sellers' labels. The existence of property rights in labels – that is, the power of holders of the rights to prevent other sellers from using holders' labels – is necessary for the benefits of labels to be enjoyed.

In view of the social value of property rights in labels, it is not surprising that the legal system allows such rights, according to trademark law. Also, trademarks are

[72] On insider trading, see Leland (1992) and Scott (1998).

[73] See Ginsburg and Shechtman (1993), R.A. Posner (1993b), and Shavell (1993c).

of potentially unlimited duration (unlike patents or copyrights), which makes sense because the rationale for their use does not wane over time. The guiding principle of trademark protection is prevention of consumer confusion, so that a new trademark that is so similar to another (Liz Clayborne and Liz Claiborne) that it would fool people would be barred, but an identical trademark might be allowed if used in a separate market. Trademarks are required to be distinctive words or signs, for otherwise normal usage would be encumbered. (If a restaurant obtained a trademark on the words "fine food", other restaurants would be limited in their ability to communicate.) On the economics of trademark law, see Landes and Posner (1987b).

4. Contracts

The private and social functions of contracts and of contract law are examined here. In Section 4.1 the basic theory of contracts is considered, in Section 4.2 production contracts (which have been the focus of a substantial literature) are analyzed, and in Section 4.3 several other types of contract are discussed.

4.1. Basic theory

4.1.1. Definitions and framework of analysis

A contract is a specification of the actions that named parties are supposed to take at various times, as a function of the conditions that then obtain. The actions usually comprise delivery of goods, performance of services, and payments of money, and the conditions include uncertain contingencies, past actions of parties, and messages sent by them.

A contract is said to be complete if the list of conditions on which the actions are based is exhaustive, that is, if the contract provides *explicitly* for all possible conditions. Otherwise, a contract will be referred to as incomplete. Typically, incomplete contracts do not include conditions that, were they easy to include, would allow both parties to be made better off in an expected sense. It should be noted that an incomplete contract may well not have literal gaps in that it will cover all conditions, at least by implication. Consider, for example, a contract stating merely that a specified price will be paid for a bushel of wheat. Although this contract is incomplete because it does not mention many contingencies that might affect the buyer or the seller of wheat, it has no gaps, as it stipulates what the parties are to do (pay a price, deliver a bushel of wheat) in all circumstances.

A contract is Pareto efficient if it is impossible to modify in a manner that raises the expected utility of both of the parties; such a contract will sometimes be referred to simply as efficient or as mutually beneficial.

Contracts are assumed to be enforced by a tribunal, which will usually be interpreted to be a state-authorized court, but it could also be another entity, such as an arbitrator or

the decision-making body of a trade association or a religious group. (Reputation and other non-legal factors may also serve to enforce contracts but will not be examined here[74].) Enforcement refers to actions taken by the tribunal when parties to the contract decide to come before it. Tribunals may impose money sanctions – so-called damages – for breach of contract or insist on *specific performance* of a contract – require parties to do what a contract specifies (for example, convey land). Tribunals may also fill gaps, settle ambiguities, and override terms in contracts.

4.1.2. Contract formation

The formation of contracts is of interest in several respects.

Search effort. Parties expend effort in finding contracting partners, and it is apparent that their search effort will not generally be socially optimal. On one hand, they might not search enough: because the surplus gained when one party locates a contract partner will generally be divided between them in bargaining, the private return to search may be less than the social return. On the other hand, parties might search more than is socially desirable because of a negative ("common pool") externality associated with discovery of a contract partner: when one party finds and contracts with a second, other parties are thereby prevented from contracting with that party[75]. Both of these externalities arise in Diamond and Maskin (1979), who examine a specific model of search and contracting. Although policies to promote or to discourage search might be desirable, one wonders whether social authorities could obtain the information needed to determine the nature of problems with search effort.

Mutual assent and legal recognition of contracts. A basic question that a tribunal must answer is at what stage of interactions between parties does a contract become legally recognized, that is, become enforceable. The general legal rule is that contracts are recognized if and only if both parties give a clear indication of assent, such as signing their names on a document. This rule obviously allows parties to make enforceable contracts when they so desire. Moreover, because the rule requires mutual assent, it protects parties against becoming legally obligated against their wishes. Thus, it prevents the formation of what would be undesirable contracts, and it means that search for contracting partners will not be chilled due to the risk of unwanted legal obligations.

However, certain legal doctrines sometimes result in parties becoming contractually bound without having given their assent; there exist cases in which a party became contractually bound when the other party with whom he was negotiating made substantial investments in anticipation of contract formation. This legal policy not only may result in undesirable contracts, it may also induce wasteful early investment as a strategy to achieve contract formation. It is true that early investment is sometimes

[74] See, for example, Bernstein (1992, 1998), Charny (1990), Greif (1998), and Klein and Leffler (1981).

[75] Compare our discussion in Section 3.5.1 of excessive incentives to search for unowned property.

efficient, but a party who wants to make early investment could attempt to advance the time of contract formation or make a preliminary contract about the matter. See Bebchuk and Ben-Shahar (2001), Craswell (1996), Katz (1996), and Wils (1993).

Offer and acceptance. Mutual assent sometimes is not simultaneous; one party will make an offer and time will pass before the other agrees. An issue that this raises is how long, and the circumstances under which, the offeror will want to be held to his offer, and whether he should be held to it. If an offeror is held to his terms, offerees will often be led to invest effort in investigating contractual opportunities. Otherwise, offerees might be extorted by offerors if the offerees expressed serious interest after investigation. The anticipation of such offeror advantage-taking would reduce offerees' incentive to engage in investigation and thus diminish mutually beneficial contract formation. Hence, it may be in offerors' and society's interests for offered terms to be enforced for some period of time. Yet offerors' circumstances may change, making it privately and socially advantageous for them to alter contract terms. On this and other issues concerning offer and acceptance, see Craswell (1996) and Katz (1990b, 1993).

Disclosure. The law may impose an obligation to disclose private information at the time of contract formation[76]. Such a legal duty is beneficial in the respect that disclosed information may be desirably employed by the buyer; suppose, for instance, that he learns from the seller that the basement of his new house leaks and thus decides not to store valuables there. However, as initially emphasized by Kronman (1978a), a disclosure obligation discourages parties from investing in acquisition of information. For example, a company might decide against conducting aerial surveys to determine the mineral-bearing potential of land if it would be required to disclose its findings to sellers of land, as sellers would then demand a price reflecting the value of the land. The social welfare consequences of the effect of a disclosure obligation on the motive to acquire information, analyzed in Shavell (1994), depend on whether the information is socially valuable or mere foreknowledge, on whether the party acquiring information is the buyer or the seller, and on inferences that would be made from silence[77].

Duress and emergency. Even if both parties have given their assent, a contract will not be recognized if it was made when one of the parties was put under undue pressure, as when he is physically or otherwise threatened by another. This legal rule has virtues similar to those of laws against theft; it reduces individuals' incentives to expend effort making threats and to defend against them.

In addition, contracts may not be legally recognized if they are made in emergency situations, such as when the owner of a ship in distress promises to pay an exorbitant amount for rescue. Nonenforcement in such situations beneficially provides victims with implicit insurance against having to pay high prices, but it also reduces incentives

[76] For discussions of various ways that asymmetry of information affects the contract terms that parties will agree to when there is no compulsory disclosure, see Ayres and Gertner (1989), Bebchuk and Shavell (1991), Spier (1992b), and Stole (1992).

[77] On inferences from silence in other contexts, see Fishman and Hagerty (1990), Grossman (1981), and Milgrom (1981), and for an empirical study of mandatory disclosure, see Mathios (2000).

for rescue (yet rescue incentives might tend to be excessive, for the same reasons that there is excessive fishing effort)[78].

4.1.3. Why contracts and their enforcement are valuable to parties

At the most general level, parties make contracts when they have a need to make plans. They want contracts enforced to ensure that promised payments are made and to prevent opportunistic behavior that otherwise might occur over the course of the contractual relationship and stymie fulfillment of their plans. There are two basic contexts in which parties make enforceable contracts.

The first is that concerning virtually any kind of financial arrangement. The necessity of contract enforcement here is transparent. For example, because borrowers would not be forced to repay loans in the absence of contract enforcement, loans would be unworkable without enforcement. In financial arrangements, there is often a party who extends credit to another for some time period, and contract enforcement prevents his credit from being appropriated, which would render the arrangements impossible. In addition, financial contracts that allocate risk would generally be made useless without enforcement because, once the risky outcome became known, one of the parties would not wish to honor the contract.

The second context in which parties make enforceable contracts involves the supply of custom or specialized goods and services – those which cannot simply be purchased on a spot market in a simultaneous exchange for money. The need for enforcement of agreements for supply of custom goods and services inheres mainly in averting what is often described as the holdup problem (discussed further in Sections 4.2.1 and 4.2.2). To illustrate, consider a buyer who wants a custom desk that would be worth $1000 to him and would cost $700 for a seller to produce. In the absence of contract enforcement, the buyer will not pay the seller in advance (for the seller could walk away with what he receives). The buyer will pay the seller only after the seller makes the desk. But at that point, the seller's production cost is sunk and he is vulnerable to holdup; the situation is that he has a desk that, being custom-made, has little or no alternative value[79]. The outcome of bargaining between him and the buyer might thus be a price lower than the seller's cost of $700; say the price is $500. If so, and the seller anticipates receiving only the $500 price, he will not produce the desk. This is true even though production and sale at a price between $700 and $1000, such as $800, would be mutually beneficial for the seller and the buyer. Enforcement of the buyer's promise to pay $800 for the desk on delivery, or of the seller's promise to produce and

[78] On rescue, see Landes and Posner (1978).

[79] Similar forms of holdup would arise in the absence of contract enforcement where parties want to convey property that already exists, such as land; for instance, a seller might worry about being held up by the buyer if he waits and forgoes a present opportunity to sell his land to a new party who makes a bid for it.

deliver the desk (if the buyer paid the price of $800 in advance), is thus desirable for the parties.

More broadly, enforcement of contracts will stimulate all manner of investments that, like the seller's expenditure on production, have specific value in a contractual relationship. Enforcement will lead buyers to train workers to use new contracted-for equipment, sellers to engage in research to reduce production costs, and so forth. In the absence of contract enforcement, there would be too little investment in these things, for, at the final stage of negotiation for performance and for payment, each side would be subject to holdup by the other, so would tend to obtain only a part of the surplus created by its investment.

The foregoing idea of contract enforcement as a cure for holdup-related underinvestment was initially stressed in the economics literature by Klein, Crawford and Alchian (1978), Grout (1984), and Williamson (1975). However, the general notion that contract enforcement is privately and socially desirable because it fosters production and trade is made (usually with little articulation) by most writers on contract law and, one supposes, has always been appreciated. See, for example, Farnsworth (1982, pp. 16–17) and Pound (1959, pp. 133–134).

4.1.4. Incomplete nature of contracts and their less-than-rigorous enforcement

Although enforceable contracts are desirable, they are observed to be substantially imperfect. They are significantly incomplete, leaving out all manner of variables and contingencies that are of potential relevance to contracting parties, and they also often fail to employ included variables in a mutually beneficial manner. Moreover, contracts are not enforced rigorously, despite the seeming strength of the reasons for contract enforcement: penalties for violation of contractual obligations are often modest, and breach is not an uncommon event.

There are three important reasons for the incompleteness of contracts. The first is the cost of writing more complete contracts. Parties may not include variables in a contract, or not in a detailed, efficient way, due to the cost of evaluating, agreeing upon, and writing terms. (In particular, parties will tend not to specify terms for low-probability events, because the expected loss from this exclusion will be minimal, whereas the cost of including the terms is borne with certainty.)

The second reason for incompleteness is that some variables (effort levels, technical production difficulties) cannot be verified by tribunals[80]. Of course, many such variables can be made verifiable (effort could be made verifiable through videotaping), but that would involve expense.

The third reason for the incompleteness of contracts is that the expected consequences of incompleteness may not be very harmful to contracting parties.

[80] The problem of unverifiability of variables is diminished by the possibility that parties can plan in their contract to use a tribunal of experts in their area, such as individuals in the same business as the contracting partners. In many industries, this practice is common.

Incompleteness may not be harmful simply because a tribunal might interpret an imperfect contract in a desirable manner. In addition, as we shall see, the prospect of having to pay damages for breach of contract may serve as an implicit substitute for more detailed terms. Furthermore, the opportunity to renegotiate a contract often furnishes a way for parties to alter terms in the light of circumstances for which contractual provisions had not been made. Finally, in some settings parties' concern for their reputation may induce them to refrain from opportunistic behavior.

That contracts are less than rigorously enforced is intimately related to their incompleteness. For incomplete contracts not to disadvantage parties, tribunals must be able to reinterpret or override imperfect contractual terms rather than always enforce these terms as written. Also, for damage measures for breach to be employed beneficially by parties, notably for parties to be able to escape from contractual obligations when performance and renegotiation are difficult, damages payments must not be excessive. Additionally, for parties to avoid bearing high risks in the form of payments that they would be induced to make when renegotiating imperfect contractual terms, the damages for breach must again not be severe. These points will be expanded in the discussion below of contract interpretation, remedies for breach, and renegotiation.

4.1.5. Interpretation of contracts

Contractual interpretation, which includes a tribunal's filling gaps, resolving ambiguities, and overriding literal language, can benefit parties by easing their drafting burdens or reducing their need to understand contractual detail[81]. For example, if it is efficient to excuse a seller from having to perform if his factory burns down, the parties need not incur the cost of specifying this exception in their contract, assuming that they can trust the tribunal to interpret their contract as if the exception were specified[82].

It may be worthwhile elaborating somewhat by viewing contract interpretation more formally, as a function that transforms the contract individuals write into the effective contract that the tribunal will enforce. Given a method of interpretation, parties will choose contracts in a constrained-efficient way. Notably, if an aspect of their contract would not be interpreted as they want, the parties would either bear the cost of writing a more explicit term that would be respected by the tribunal, or else they would not

[81] On various aspects of contract interpretation, see, for example, Ayres and Gertner (1989), Hadfield (1994), and Schwartz (1992).

[82] Another example, where it may be efficient for a tribunal to override particular terms that appear in contracts, is when a seller offers only a detailed, fine-print contract in conjunction with the sale of an inexpensive good or service. Because it would be irrational for consumers to read such contracts, sellers would have incentives to include inefficient, one-sided terms if such terms would be enforced. See Katz (1990c). The extent to which such contracts will be problematic will depend on the fraction of consumers who are informed about contract terms and shop among competing sellers. See Schwartz and Wilde (1979).

bear the cost of writing the more explicit term and accept the expected loss from having a less than efficient term. The best method of contract interpretation will take this reaction of contracting parties into account and can be regarded as implicitly minimizing the sum of the costs the parties bear in writing contracts, the losses resulting from inefficient enforcement, and adjudication costs[83].

4.1.6. Damage measures for breach of contract

When parties breach a contract, they often have to pay damages in consequence. The damage measure, the formula governing what they should pay, can be determined by the tribunal or it can be stipulated in advance by the parties to the contract[84]. One would expect parties to specify their own damage measure when it would better serve their purposes than the measure the tribunal would employ, and otherwise to allow the tribunal to select the damage measure. In either case, we now examine the functioning and utility of damage measures to contracting parties (assuming here that there is no renegotiation of contracts).

Clearly, the prospect of payment of damages is an incentive to perform contractual obligations, and thus generally promotes enforcement of contracts and the goals of the parties, as discussed in Section 4.1.3. As emphasized in Section 4.1.4, however, damages for breach in fact are not chosen to be so high that they virtually guarantee performance of contracts as written. Under the commonly employed measure of *expectation damages,* damages equal the amount that compensates the victim of breach for his losses.

Why are damages not chosen to be so high as to guarantee performance? An important explanation is that parties do not always want performance of the less-than-complete contracts that they write. For example, suppose that a contract is very incomplete: it merely states, "The seller will produce a custom desk for the buyer and receive full payment of $800 in advance." The buyer and the seller do not really want the desk always to be produced. It is readily shown that, had they made a Pareto-efficient complete contract, they would have specified that there should be performance if and only if the production cost is less than the $1000 value of the desk to the buyer. (For instance, in a complete contract, they would have jointly decided against a contractual term specifying performance when the production cost is $2000, for the

[83] The determination of the optimal method of interpretation may involve subtleties. For example, according to the optimal method, a term might not be interpreted in the way that is best in the majority of transactions. Suppose that term A is best in the majority of transactions and that the parties to these transactions can include A explicitly, at little cost on a per-contract basis, because they are repeat players. Suppose that term B is best only in the minority of transactions, but that for the parties to these transactions to include B explicitly will not be cheap on a per-contract basis because they are not repeat players. Then the optimal method of interpretation would make B the default term even though it is best in only a minority of transactions.

[84] A contractual provision that states a particular amount of damages is referred to as a *liquidated damages* clause.

seller would have been willing to reduce the contract price sufficiently to induce the buyer to strike the term.) Now if the incomplete contract calling for the desk always to be produced is enforced by the expectation measure of damages of $1000, the seller will behave exactly as he would have under the Pareto-efficient complete contract, that is, he will perform if and only if the production cost is less than $1000. Higher damages than the expectation measure might induce performance when it is inefficient, and lower damages might lead to breach when that is inefficient. Indeed, for this reason, the parties would often agree to choose the expectation measure over other measures of damages.

This understanding of damage measures as a device to induce the behavior that the parties would have specified in more complete contracts sheds light on the notion held by some legal commentators and philosophers that contract breach is immoral, that it constitutes the breaking of a promise. That belief is often incorrect, it is submitted, and might fairly be considered to be the opposite of the truth. The view that a contract breach is the breaking of a promise overlooks the point that the contract that is breached is generally an incomplete contract, and that the "breach" constitutes behavior that the parties truly want and would have provided for in a complete contract. In the example of the simple incomplete contract calling for a desk to be produced, the seller who finds that his production cost would be $2000 will commit breach under the expectation measure. But in so doing, he will be acting precisely as would have been set out in a Pareto-efficient complete contract, and it is that contract which is best regarded as the promise between the parties that ought to be kept.

The point that a moderate damage measure, and in particular the expectation measure, is desirable because it induces performance if and only if the cost of performance is relatively low was apparently first clearly stated (informally) in R.A. Posner (1972), who emphasized the social efficiency of the measure. Shavell (1980b) formally demonstrated this and also stressed the mutual desirability of the expectation measure for contracting parties and its role as a substitute for more complete contracts[85].

Several more comments should be made about damage measures and incentives. First, damage measures influence the motive of contracting parties to make reliance investments (so called because the investments are made relying on contract performance). Reliance investments are illustrated by the earlier-noted instance of a buyer training workers to use a contracted-for machine or by advertising the contracted-for appearance of an entertainer. Under the expectation measure, there is a tendency for reliance investment to exceed the Pareto efficient level: the buyer will treat an investment like advertising as one with a sure payoff – either he will receive performance or receive expectation damages, a form of insurance – whereas the actual

[85] Two other writers, Birmingham (1970) and Barton (1972), adumbrate these points, although the meaning of their articles is at times obscure. See also Diamond and Maskin (1979), who consider damage measures in analyzing search behavior.

return to investment is uncertain, due to the possibility of breach (advertising will be a waste if the entertainer does not appear). This tendency toward overreliance due to the receipt of contract damages was initially noted in Shavell (1980b), and stands in contrast to the problem of inadequate reliance investment associated with lack of contract enforcement. The issue of reliance investments has been elaborately analyzed, as will be described in Section 4.2.2.

A second comment is that the value of damage measures as an incentive toward efficient performance would not exist if renegotiation of contracts in problematic contingencies would always result in efficient performance. But, as will be discussed below, it seems plausible that renegotiation would not always result in efficiency.

An important function of damage measures that is quite distinct from their incentive role concerns risk-spreading and compensation. Notably, because the expectation measure compensates the victim of a breach, the measure might be mutually desirable as a form of insurance if the victim is risk averse. However, the prospect of having to pay damages also constitutes a risk for a party who might commit breach (such as a seller whose costs suddenly rise), and he might be risk averse as well. The latter consideration may lead parties to want to lower damages (see Polinsky 1983) or to avoid use of damages as an incentive device, by writing more detailed contracts (for instance, the parties could go to the expense of specifying in the contract that a seller can be excused from performance when his costs are high)[86]. A full consideration of damage measures and efficient risk allocation would also take into account whether the risk that a party bears is detrimental or beneficial[87], whether the risk is monetary or non-monetary[88], and whether the parties can obtain insurance.

4.1.7. Specific performance as a remedy for breach

As observed at the outset, an alternative to use of a damage measure for breach of contract is specific performance: requiring a party to satisfy his contractual obligation[89]. Specific performance can be accomplished with a sufficiently high threat or by exercise of the state's police powers, such as by a sheriff removing a person from the land that he promised to convey. (Note that if a monetary penalty can be employed

[86] When parties do not so specify in advance, certain legal doctrines may serve this function. See Joskow (1977), R.A. Posner and Rosenfield (1977), and Sykes (1990).

[87] For example, if a party wants to breach because he has a superior opportunity, optimal damages might be higher, although adjusting damages in the case of beneficial risks is not likely to matter as much on risk-bearing grounds.

[88] For example, if the victim's loss is non-monetary, such as the loss due to failure of musicians to appear at a wedding, financial compensation in the form of damages may not constitute an optimal form of insurance. See Section 2.4.2.

[89] Some economists have employed the term "specific performance" in an unconventional sense, to refer to enforcement of all provisions in a contract, including any damage measure named in it. Thus, they would say that a contract is specifically performed when the parties name expectation damages in their contract and parties who breach are thus required to pay these damages.

to induce performance, then specific performance is equivalent to a damage measure with a high level of damages.)

It is apparent from what has been said about incomplete contracts and damage measures that parties should not want specific performance of many contracts that they write, for they do not wish their incomplete contracts always to be performed. It is therefore not surprising that, in fact, specific performance is not used as the remedy for breach for most contracts for production of goods and for provision of services. Additionally, it may be observed that specific performance might be peculiarly difficult to enforce in these contexts because of problems in monitoring and controlling parties' effort levels and the quality of production.

However, specific performance does have advantages for parties in certain contexts, such as in contracts for the transfer of things that already exist, like land, and specific performance is the usual legal remedy for sellers' breaches of contracts for the sale of land. This point is discussed briefly below, in Section 4.3.1. On specific performance and its general comparison to damage remedies, see Bishop (1985), Kronman (1978b), Schwartz (1979), Shavell (1984b), and Ulen (1984). (Specific performance also is examined in some of the articles on production contracts cited in Section 4.2.2.)

4.1.8. Renegotiation of contracts

Parties often have the opportunity to renegotiate their contracts when problems arise. Indeed, the assumption that they will do this has appeal because, having made an initial contract, the parties know of each other's existence and of many particulars of the contractual situation. For this reason, much of the economics literature on contracts assumes that renegotiation always occurs when inefficiency would otherwise result; see, for example, Hart (1987), Hart and Holmström (1987), and Rogerson (1984).

Nevertheless, in many circumstances contracts will not be renegotiated because parties are not in contact with each other when difficulties are experienced and one party would benefit from acting quickly. A problem may occur during the course of production and the producer may have to decide on the spot whether to abort the process or proceed at greater cost. Or a new bid may be heard and have to be immediately answered. Furthermore, even if the parties are in contact with one another, asymmetric information may lead to breakdowns in renegotiation.

In any event, let us assume that successful renegotiation tends to occur and consider how it affects the welfare of contracting parties. Plainly, renegotiation often allows parties to avert Pareto-inefficient breach decisions. For example, if damages exceeding the expectation measure or specific performance were the remedy for breach, a seller might be led to perform when his production cost exceeds the value of performance to the buyer. To avoid this inefficient outcome, the seller might pay the buyer to release him from his obligation to perform. That renegotiation may result in performance if and only if it is efficient means, as we noted, that damage measures for breach are not necessary to accomplish this, and also helps to explain why contracts lack detail.

But even if renegotiation tends to occur, it may represent only a partial substitute for explicit contractual terms or for appropriate damage measures for breach. One reason (see Section 4.2.3) is that renegotiation cannot affect actions that are taken before the time of renegotiation, which influence the likelihood of nonperformance; renegotiation can only affect future decisions about breach. Another reason involves the allocation of risk-bearing. Consider, for instance, the substantial risks borne by a producer who may have to purchase a release from an obligation to perform when his production costs would be extremely high. Such risks could be mitigated by use of a clause excusing him from performance or by a damage measure such as expectation.

Additionally, the prospect of renegotiation affects the incentives of parties to invest in the contractual relationship. A party's level of reliance investment will be inefficient if renegotiation results in the extraction of part of the surplus that the party's reliance investment creates. Yet renegotiation is influenced by, among other elements, the damage measure that applies for breach, and if the damage measure is appropriately chosen, the damage measure together with renegotiation may, in principle, spur desirable reliance investment; see Section 4.2.2.

One presumes that the ability to renegotiate is usually desirable for contracting parties, because it allows them to improve their situation when difficulties arise and to write simpler contracts than otherwise. Thus, we would expect that parties will want their renegotiated contracts enforced, and the law generally does enforce renegotiated contracts. However, the ability to renegotiate can also work to the detriment of parties because they might thereby be prevented from committing themselves to particular outcomes in their initial contract. See Jolls (1997) and the literature cited therein, especially Fudenberg and Tirole (1990). Nevertheless, the law usually prevents parties from binding themselves not to renegotiate, even though that could in theory be done[90].

4.1.9. Legal overriding of contracts

A basic rationale for legislative or judicial overriding of contracts is the existence of externalities. Contracts that are likely to harm third parties are often not enforced, for example, agreements to commit crimes, price-fixing compacts, liability insurance policies against fines, and certain simple sales contracts (such as for machine guns)[91].

[90] It is true that parties will not usually be able to bind themselves against engaging in renegotiation, for they could ordinarily renegotiate in secret. However, as Jolls (1997) observes, one of the parties will usually prefer that the original contract be enforced, so that if the courts stand ready to enforce the original contract, renegotiation cannot result in a new contract. For example, in the standard principal–agent contract, after the agent exerts effort, the principal and the agent will have an incentive to arrange for the agent to be paid a constant amount. But if this were contemplated, then after the output is realized, the agent would have an incentive to assert the original contract if his pay would be higher according to it, and the principal would assert the original contract if he could pay less under it.

[91] See also E.A. Posner (1995), who suggests that such contractual limits as usury laws, which constrain consumers' ability to borrow, might be justified by a type of externality: when high-risk borrowers fail, they may become eligible for social welfare programs, imposing costs on taxpayers.

Another general rationale for nonenforcement of contracts is to prevent a loss in welfare to one or both of the parties to contracts. This concern may motivate nonenforcement when a party is incompetent, lacks relevant information, or is in an emergency situation (see Section 4.1.2). The rationale also applies in the context of contract interpretation by tribunals; as discussed in Section 4.1.5, contract interpretation may amount to overriding terms of contracts, and this may promote the welfare of contracting parties by allowing them to write simpler contracts[92].

Additionally, contracts sometimes are not enforced because they involve the sale of things said to be inalienable, such as human organs, babies, and voting rights. In many of these cases, the inalienability justification for lack of enforcement can be recognized as involving externalities or the welfare of the contracting parties[93].

4.2. Production contracts

In this section, the literature on production contracts is discussed. The first case considered is that where symmetrically informed, risk-neutral parties enter into contracts, and the only variables of concern are the value of performance and production cost. Then the case where parties make reliance investments to raise the value of the contract during the contract period is examined. Finally, several other issues, including risk-bearing and asymmetric information, are reviewed. Throughout, when remedies for breach are discussed, one can imagine them to be chosen either by the parties or by the courts.

4.2.1. Value of performance and production cost

Assume that a risk-neutral buyer and a risk-neutral seller have met; the seller faces uncertain production cost c, which he will learn before he decides whether to produce; v is the certain value of performance to the buyer; and the parties are symmetrically informed. The Pareto efficient outcome is for the seller to produce if and only if $c < v$. (That is, in a complete contract, with terms for all contingencies, performance would be required if and only if $c < v$; a change in the contract price would compensate a party for agreeing to alter a term from any initially considered contract under which performance does not occur if and only if $c < v$.)

In the absence of contract enforcement, then (amplifying on Section 4.1.3) there would be too little production because the buyer would only pay the seller for actual delivery of the good and cannot guarantee the price. In particular, supposing that the seller would obtain a fraction α of the surplus from a transaction (α reflects bargaining strength), he would obtain a price of αv. (After the seller produces the good, the surplus

[92] Also, at least in theory, nonenforcement of contracts might also be beneficial to parties where they would be led to include terms constituting wasteful signals of unobservable characteristics. See Aghion and Hermalin (1990).

[93] See generally Rose-Ackerman (1985) and Trebilcock (1993).

from the transaction would be v, presuming for simplicity that the custom good has no alternative value for the seller.) Thus, the seller would decide to produce only when $c < \alpha v$, rather than whenever $c < v$.

Suppose now that there is contract enforcement and that the parties are not able to renegotiate before the seller decides whether to produce (an assumption that is relaxed below). If c is verifiable by the tribunal, the parties could write a complete contract specifying performance if and only if $c < v$. The parties would want a damage measure d for breach of this contract to be sufficiently high to induce performance when $c < v$, and thus any d exceeding c would work.

If c is not verifiable, the parties are able to write an incomplete contract specifying "The seller shall deliver the good to the buyer, who will pay price p at the outset", accompanied by damages d for seller breach. Under such a contract, the seller will perform when $c < d$ and will commit breach otherwise[94]. If the expectation measure is employed, that is, $d = v$, the seller will perform if and only if $c < v$, so that performance will be efficient[95]. If damages d exceed v, there will be excessive performance, as there will be if there is specific performance. If d is less than v, there will be too little performance. The points of these paragraphs were, as noted, emphasized in R.A. Posner (1972) and Shavell (1980b).

If, instead, it is assumed that the buyer and the seller can renegotiate their contract after c becomes known but before the seller decides whether to produce, then, given symmetric information, it is natural to suppose that there will always be Pareto efficient performance, regardless of d.

Let us also note that if the buyer's value v is uncertain as well as the seller's production cost c, the major difference in the outcome is that, since v cannot be prescribed as damages in the contract, v must be verifiable for the expectation measure $d = v$ to be applied by the tribunal (c still need not be verifiable)[96].

[94] Because we assume that the price p is paid at the outset, the seller faces cost c if he performs and will compare it to damages of d that he would have to pay if he breaches. If the price were to be paid only at the time of performance, then the seller would perform if and only if $c - p < d$. Hence, the performance that is induced under d if the price is paid at the outset will be achieved under $d' = d - p$ if the price is paid only at performance.

[95] A related issue concerns post-breach mitigation behavior of the buyer: efficiency requires that if there is a breach, the buyer should mitigate the consequences of breach by searching for alternative suppliers and the like. Let z be mitigation expenditure of the buyer to raise his post-breach alternative value, say $w(z)$. Efficiency requires the buyer to choose z to maximize $w(z) - z$; let z^* be the optimal value of z. If y is the gross value of seller performance to the buyer, then we can define v, the net value of performance, as $v = y - (w(z^*) - z^*)$. Thus, expectation damages for breach should equal this v, not the gross value y. And if damages equal v, then the buyer will choose z^* if he is the victim of a breach, and the net value of performance will actually be v. On this issue of mitigation of the consequences of breach, see, for example, Wittman (1981).

[96] However, if c is verifiable and v is not, Pareto-efficient performance can be achieved by constructing the contract so that the *buyer* will commit breach by refusing to pay for performance when performance would be inefficient. Specifically, let the price p be paid at performance, and let damages for buyer breach be

4.2.2. Reliance investment during the contract period

Now assume that parties can make investments during the period of the contract that affect its value v or the production cost c. Such investments are, as noted, sometimes called reliance investments, since they are made in anticipation of contractual performance. We will begin with the case in which just one party invests before discussing the case where both sides invest.

Buyer makes reliance investment and seller's costs are uncertain. Suppose that one party to the contract invests, for concreteness the buyer, and that the other party faces uncertainty[97]. Specifically, let r be the buyer's reliance investment (training of workers to use a contracted-for machine) and let $v(r)$ be the value of performance given r, where v is increasing in r. The buyer chooses r before the seller learns c and decides about producing. The Pareto-efficient decision of the seller is to produce when $c < v(r)$, and the efficient decision of the buyer is therefore to choose r to maximize

$$\int_0^{v(r)} (v(r) - c)\, g(c)\, \mathrm{d}c - r,$$

where g is the density of c. Thus, the optimal r, denoted r^*, is determined by $v'(r)\, G(v(r)) = 1$, where G is the cumulative distribution of c. The point to note here is that the marginal return to reliance investment is only a contingent return, for the investment pays off only with probability $G(v(r))$, when $c < v(r)$ (when production turns out to be efficient).

In the absence of contract enforcement, there will be too little production, as before; it will occur only when $c < \alpha v(r)$. But now, in addition, the buyer will choose an incorrect value of r because he will only obtain a fraction $1 - \alpha$ of the value created by investment[98].

Assume next that there is contract enforcement and that the parties do not renegotiate before the seller's production decision (we relax this assumption below). This is the setting analyzed in Shavell (1980b), who first studied reliance investment. If c and r

$d = p - c$, the seller's profits. Then the buyer will breach and refuse performance whenever $v - p < -(p - c)$, or when $v < c$. (If, as is realistic, it is assumed that $p - c$ cannot be negative, then the parties can choose p high enough that it always exceeds c (assuming c is bounded), with the buyer being compensated for the high p through an up-front rebate.) The parties' ability to determine who will make the breach decision, as described here, is emphasized in Edlin (1996).

[97] We comment in note 99 below on another case of reliance investment: where the party who chooses the reliance investment is the same party who faces uncertainty, such as where the seller chooses r to lower his production cost and faces uncertainty about his production cost.

[98] Specifically, he will choose r to maximize $(1 - \alpha)\, v(r)\, G(\alpha v(r)) - r$, so the first-order condition determining r is $(1 - \alpha)\, v'(r)\, G(\alpha v(r)) + (1 - \alpha)\, v(r)\, \alpha v'(r)\, g(\alpha v(r)) = 1$, or $(1 - \alpha)\, v'(r)[G(\alpha v(r)) + \alpha v(r)\, g(\alpha v(r))] = 1$. Although one might expect r to be less than r^*, it is apparent from the first-order condition that there is a possibility that the r chosen would exceed r^*. The reason is that increasing r raises the probability that the buyer will obtain performance from the seller.

are verifiable by the tribunal, the parties can write a contract specifying efficient performance (when $c < v$) and also specifying r^*; again, they would want the contract enforced by a damage measure high enough to ensure performance, and any such measure of damages would serve their purposes.

Now assume that c and r are not verifiable, that the parties write a simple contract specifying "The buyer will pay price p at the outset and the seller will deliver the good to him", and consider what occurs under different damage measures. If the expectation measure is employed, that is, $d = v(r)$, the seller will perform when $c < v(r)$, so that performance will be efficient. However, as the buyer will always receive $v(r)$ (either he obtains performance, worth $v(r)$, or damages of that amount), he will choose r to maximize $v(r) - r$. Consequently, the buyer will select an inefficiently high r; the problem is that the buyer does not take into account that investment does not have any value when performance does not occur[99]. Under a sophisticated expectation measure based on efficient investment, namely $d = v(r^*)$, however, investment as well as performance can be shown to be efficient[100].

Another damage measure that has been examined is known as *reliance damages*, according to which the buyer would receive a return of his initial payment p plus his reliance investment r if the seller breaches. Under this measure, if there is a breach,

[99] We observe that the problem of an inefficiently high r does not arise under the expectation measure where the seller makes the reliance investment to lower his production cost and also faces uncertainty about it. Specifically, suppose that production cost is $c(r, \theta)$, where θ is an uncertain state of nature, $c_\theta > 0$, and $c_r < 0$. In this case, the seller will choose the efficient r; the explanation in essence is that the seller obtains the benefit of his reliance only when there is performance. The efficient r is that maximizing

$$\int_0^{\theta(v,r)} (v - c(r, \theta)) g(\theta)\, d\theta - r,$$

where $\theta(v, r)$ is the θ such that $c(r, \theta) = v$. Thus, r^* is determined by $-\int c_r(r, \theta)) g(\theta)\, d\theta = 1$. Now under the expectation measure, the seller will perform when $c < v$ and pay v otherwise. Thus, the seller chooses r to maximize

$$p - \int_0^{\theta(v,r)} c(r, \theta)\, g(\theta)\, d\theta - v(1 - G(\theta(v, r))) - r,$$

and differentiation of this yields the same condition as that which determines r^*. This point is noted in Shavell (1980b).

[100] If $d = v(r^*)$, the seller will perform when $c < v(r^*)$, so the buyer will maximize $v(r)\, G(v(r^*)) + v(r^*)(1 - G(v(r^*))) - r$. Accordingly, r will be determined by $v'(r)\, G(v(r^*)) = 1$, and this condition is clearly satisfied at r^*. The explanation is that the buyer's choice of r affects his return only when he obtains performance. Hence, r^* will be chosen and performance will also be efficient. This point was first mentioned by Cooter (1985). (Observe that the tribunal does not need to observe r to enforce $d = v(r^*)$, as the parties can name $v(r^*)$ in the contract.) The analysis would change, however, if the buyer does not know $G(\cdot)$. See Craswell (1988).

the buyer will be placed in the position he would have enjoyed had he not invested and made the contract. It can be shown that, under the reliance measure, investment would be even more excessive than under the expectation measure, and there would be too little performance. (Note, however, that to apply the reliance measure, courts must be able to verify investment r, and that if this is so, r^* could be achieved simply by the parties naming it in their contract.) Finally, under specific performance, there is excessive performance, but r is chosen optimally given that level of performance (because performance always occurs).

Next assume that the parties do renegotiate after the reliance investment is made and before the seller decides about production, so that, assuming symmetric information, there will always be efficient performance. This version of the model of production contracts was originally studied by Rogerson (1984). Here, damage remedies may influence investment through their effect on the outcome of renegotiation. To illustrate, consider what would occur under specific performance. Under this remedy, as suggested earlier, there will be renegotiation in which the seller pays the buyer to be allowed not to perform whenever $c > v(r)$, since then performance would be inefficient. In particular, the assumption is that the seller would pay the buyer $v(r) + (1-\alpha)(c - v(r))$ to be allowed not to perform; for $v(r)$ is needed to compensate the buyer for not receiving performance, $1-\alpha$ is the buyer's share of the surplus from renegotiation, and $c - v(r)$ is that surplus. Anticipating this, the buyer can be shown to choose an r exceeding the efficient level[101]. The features of the results about reliance investment in the case with renegotiation are very close to those where there is no renegotiation. Indeed, they are identical under the expectation measure, essentially because there is no renegotiation under the expectation measure; thus, with $d = v(r)$, investment will be excessive because the buyer will always be compensated for his investment. Furthermore, under the sophisticated expectation measure based on efficient investment, $d = v(r^*)$, investment will be efficient. See Spier and Whinston (1995).

Both parties make reliance investments and both the value of performance and production costs are uncertain. Here let $v = v(r,\theta)$ and $c = c(s,\theta)$, where s is reliance investment of the seller and θ is the state of nature; s lowers c given θ. In this more general situation, what occurs can be understood in many respects by analogy to the case just discussed. For example, under the expectation measure, investment will tend to be excessive for both parties, but performance will be efficient.

Much recent literature, beginning with Hart and Moore (1988), has focused on this general situation, assuming that parties can renegotiate after reliance investments are

[101] In general, the buyer will choose an r in between the excessive level he would choose under the expectation measure (determined by $v'(r) = 1$) and r^*. This can be explained as follows. If the buyer's fraction of surplus is 0, he will receive $v(r)$ whether or not there is renegotiation, so his situation will be the same as under the expectation measure. If the buyer's fraction of surplus is 1, he will clearly choose r^*. This suggests what can be shown, that if the buyer's fraction of surplus is positive and less than 1, he will choose an r exceeding r^* and less than the r chosen under the expectation measure.

made and θ is revealed, and that they will always then agree on efficient production decisions because information is symmetric. The literature in question, furthermore, usually supposes that none of the variables (costs, values of performance, reliance investments) are verifiable by the tribunal. Thus, a contract can depend only on what is recorded in it, certain subsequent communications between the parties, whether there has been performance, and, if not, who committed breach.

Of note are a number of results establishing the existence of contracts that will produce efficient outcomes, that is, both parties choosing efficient levels of reliance investment (performance will always be efficient). Aghion, Dewatripont and Rey (1994) and Chung (1991) demonstrate the efficiency result using a contract in which one party is effectively given the right to make a single take-it-or-leave-it offer to the other in renegotiation. It is evident that this party will invest efficiently, as he can extract in bargaining the full marginal return from his investment. For instance, if the buyer has the right to make an offer and is paying the seller to perform, he will pay only the minimum needed to induce the seller to do so, and will obtain any increase in value $v(r,\theta)$ due to his having chosen a higher r. Less apparent is how the other party is given an incentive to invest efficiently; that is accomplished by properly choosing the quantity of the good or the probability of delivery. (For instance, if the named quantity of the good is chosen to be higher than is likely to be efficient, the buyer will usually pay the seller to agree to lower the quantity. The amount the buyer will pay must compensate the seller for the profits he would have made at that higher contracted-for quantity. But the profits the seller would have made at that quantity will depend on his investment in lowering production costs – thereby giving the seller an incentive to invest in lowering his production costs, and an incentive that is greater the higher the contracted-for quantity.) Also, Noldeke and Schmidt (1995) establish that a simple option contract will induce efficient investments for reasons that are closely related to those just reviewed. Additionally, Edlin and Reichelstein (1996) and Hermalin and Katz (1993) adduce contracts leading to efficiency under somewhat different conditions, and Rogerson (1992) shows that efficiency can be achieved under wide circumstances, but assuming that parties can commit not to renegotiate their contracts.

Cooperative reliance investments. It has been assumed above that a reliance investment benefits directly only the party who makes it. Another possibility is that a reliance investment benefits the other party to the contract; importantly, suppose that a seller's investment raises product quality and, thus, value for the buyer. Such cooperative reliance investment is studied in Che and Chung (1999)[102]. As they emphasize, when cooperative investment cannot be verified by courts, then under the expectation measure, there will be too little investment (in contrast to the usual case under the expectation measure, where investment is excessive). Indeed, there will be no investment if the seller who makes a cooperative investment will not benefit directly

[102] Cooperative investment is also studied in MacLeod and Malcomson (1993).

or in damages he receives in the event of breach. Moreover, there is no contract that will result in efficient cooperative investment (again in contrast to the usual case); this point is stressed in Che and Hausch (1999), who also demonstrate that contracting offers no advantage over no contracting in a wide set of circumstances[103].

4.2.3. Further considerations

Risk-bearing. We have not discussed in this section on production contracts the allocation of risk among possibly risk-averse contracting parties, about which several comments should be made. First, if all variables are verifiable by a tribunal, the presence of risk-averse parties does not affect when it is Pareto efficient to perform; it continues to be efficient to perform if and only if $c < v$. However, efficiency requires that the resulting risk be allocated appropriately; for instance, if the seller is risk averse and the buyer risk neutral, the seller would be insured against fluctuations in c by the buyer's paying him c plus a constant. In addition, the level of efficient reliance investment will generally be affected by considerations of risk-bearing.

Second, when variables of relevance are not verifiable, then damage measures and other mechanisms that may be employed to induce efficient behavior when parties are risk neutral have to be reconsidered. For instance, the expectation measure imposes risk on the party who might breach and pay these damages; if that party is risk averse, the expectation measure would become less attractive relative to lower measures of damages. Furthermore, as we earlier noted, renegotiation does not generally lead to efficient risk bearing, even though it may lead to efficient performance.

Asymmetric information. Another factor about production contracts that we have not examined is asymmetric information between the parties. When parties are asymmetrically informed, renegotiation of contracts might not be successful, so that it becomes more important that the initial contract induces efficiency. Hermalin and Katz (1993) show that efficiency can be achieved under certain types of asymmetry of information using a relatively complicated mechanism in the contract.

New entrants. We have not examined the possibility that new buyers would appear and bid for the seller's good (a similar possibility is that new sellers would appear and make offers to the buyer). In this regard, it should be noted that it is Pareto efficient for the initial contracting parties that a sale be made to a new buyer if and only if his bid exceeds the contract buyer's valuation. Moreover, the contracting parties will want to maximize the amount that they can extract from a new buyer if he purchases the good. This observation raises the possibility that the buyer and the seller may wish to set damages for seller breach at a high level in order to induce a new party to bid more (which he would have to do to make it in the seller's interest to commit breach). Such

[103] For recent attempts to provide a unified theoretical framework for the various problems concerning reliance discussed in Section 4.2.2, see Maskin and Moore (1999) and Segal and Whinston (forthcoming).

an incentive of contracting parties to set damages at high levels can, though, result in too little breach and sale to new parties; thus, at least in principle, the incentive in question is a ground for tribunals not to enforce the high damage level specified by the contracting parties. This point was first made in Diamond and Maskin (1979) and has been refined in a number of articles; see Aghion and Bolton (1987) and Chung (1992). However, Spier and Whinston (1995) observe that three-way renegotiation would seem to vitiate the advantage to the contracting parties of setting high damages. Yet they emphasize another reason (concerning induced reliance investment) that the parties will, after all, benefit from setting high damages.

Precautions and probabilistic breach. It has been supposed throughout that breach occurs when a party decides not to perform, but often breach does not occur in this way: rather a party chooses a level of precaution that affects the likelihood of performance, and a random factor then determines whether breach or performance results. For example, a shipper's care in packing dishes affects the likelihood that they will arrive unbroken, and a chance event (a jolt) determines whether they arrive broken or unbroken. In this setting, the conclusions reached about damage measures in the absence of renegotiation continue to apply: the expectation measure results in efficient precautions, the buyer's reliance investment is excessive, and so forth. The very issue of renegotiation is made moot because the precautions are chosen before breach might occur (if the dishes arrive broken, it is too late for renegotiation). See Bebchuk and Png (1999), Cooter (1985), Craswell (1988), and Kornhauser (1983).

4.3. Other types of contract

4.3.1. Contracts for transfer of possession

A different contractual context from production is where something that already exists is to be conveyed to a buyer. Examples include contracts for transfer of real estate, goods in inventory, and durable goods. Here a major uncertainty of interest concerns bids by new parties. With regard to these bids, the points just discussed concerning new entrants apply; the parties would like for there to be a sale to a new buyer when he will pay more than the contract buyer's valuation, and so forth.

It is of interest to explore why contracting parties often adopt specific performance as the remedy for breach of contracts for transfer of possession, even though damage measures are commonly employed for other types of contract. Initially, suppose that the contract buyer and the contract seller have equal access to bids from new parties. Then the buyer's always receiving the good does not result in any loss of opportunity to sell to a new party willing to bid a high amount; that is, specific performance does not suffer from any clear disadvantage relative to damage measures that would allow the seller to breach and sell to a new party. Moreover, specific performance offers the parties an advantage over damage measures. Namely, because under specific performance it will always be the buyer who will be bargaining with a new party, the good will never be sold to a new party bidding less than the buyer's valuation. In contrast, such a sale

could occur if the seller might pay damages, commit breach, and bargain with a new party (suppose that bargaining does not involve the contract buyer as well). And, after such a sale, the buyer would have to obtain the good through repurchase from the new party, but in general this will be at a higher price than the seller obtained – meaning that some of the surplus would be shared with the new party. (Although the contracting parties would be worse off, if the buyer repurchases at a higher price, society would not be worse off as the good would still be allocated to the user who places the highest value on it.) See Shavell (1984b) and Bishop (1985).

The foregoing advantage of specific performance in preventing inefficient sales to new parties is clearly reduced if the buyer does not have equal access to bids from new parties (suppose that the seller is a dealer and the buyer is not). Also, the use of specific performance might increase transaction costs, if the new party purchases after delivery of the good to the buyer.

Notice too that some of the disadvantages of specific performance in the production context are less significant in the present context of transfer of possession. In production contracts, specific performance imposes a possibly large risk of loss on sellers whose production costs might be very high; here, specific performance only reallocates a beneficial risk (of a sale at a high price) from seller to buyer. In addition, enforcement of specific performance in the context of contracts for transfer of possession is often easier than in the production context, where enforcement might involve policing the quality of production or services.

4.3.2. Donative contracts

An important category of contractual arrangement is donative, concerning gifts. Assuming that the motivation for gifts is altruism[104], a basic question is why a donor would want to defer his gift rather than make it immediately (in which case no contract would be required). The answers include the possibilities that the donor may face liquidity constraints and that he may wish to wait for resolution of uncertainties concerning, among other factors, his own needs and future income and the donee's needs, future income, and character traits.

Given that a donor does desire to defer making a gift, would he want to make a contract that would in some way bind him? The disadvantage of so doing is that it may not be feasible for him to limit as he wishes the conditions under which he makes the gift (due to the costs of specifying these conditions and to the problems that courts would have in verifying them). The donor's principal benefit from entering into a contract is that it may induce the donee to engage in reliance activities that will increase the value of the gift to the donee (a high-school student might study more if he anticipates a gift that will finance his college education). Such reliance activities

[104] Other motives for gift giving exist (such as obtaining utility from expressions of gratitude from donees); some have similar implications to those of altruism.

will in turn inure to the benefit of the donor because of his altruism. However, if the donee knows about the altruism of the donor, a contract may not be necessary to induce donee reliance activity; if so, a contract would be disadvantageous for the donor. On these issues, see Goetz and Scott (1980) and Shavell (1991a); and see also R.A. Posner (1977b) and E.A. Posner (1997).

4.3.3. Additional types of contract

In this section, mention has not been made of many additional types of contract, including principal–agent contracts, even though they have been studied, often intensively, in the economics literature. The omission of such contracts from consideration here is explained in part by convention (by what is and is not considered to be a law and economics topic) and in part by the relative inattention that has been paid to contract enforcement.

5. Litigation

In this section, we consider civil litigation, that is, the bringing of lawsuits by private actors to enforce their rights in the areas of civil law that we have just discussed. Until now, we have largely assumed that the operation of the legal system is frictionless, in the sense that the bringing and adjudication of lawsuits is without cost. We now analyze the implications of the expense involved in the operation of the legal system.

We begin with what may be called the basic theory of litigation: the choice of a party who has suffered a loss whether to sue; if suit is brought, the choice of the litigants whether to settle with each other or instead go to trial; and the choice of litigants, before or during trial, of how much to expend on litigation. Then we discuss various extensions to the basic theory of litigation, including nuisance suits, the shifting of legal fees, lawyers and agency problems in litigation, and legal discovery. We subsequently consider the provision of legal advice, the appeals process, alternative dispute resolution, and the formulation of legal rules.

5.1. Suit

5.1.1. Private incentive to sue

As a general matter, the plaintiff will sue when the cost of suit c_P is less than the expected benefits from suit. The expected benefits from suit incorporate potential settlements and trial outcomes, but in this section we usually assume for simplicity that if suit is brought, the plaintiff obtains as a judgment a certain amount h equal to harm suffered. Thus the plaintiff will sue if and only if his litigation cost, c_P, is less than h. (Obviously, if there is only a probability p of winning this amount, the plaintiff, if risk neutral, would sue if and only if $c_P < ph$; and if the plaintiff is risk

averse, he would be less likely to sue.) The effect on the private incentive to sue of many variations in the legal environment is straightforward to identify, as we will note below.

5.1.2. Socially optimal suit versus the private incentive to sue

The private incentive to bring suit is fundamentally misaligned with the socially optimal incentive to do so, given the social costs and social benefits of suit. The deviation between the privately motivated and socially appropriate level of suit could be in either direction. The general reasons for these conclusions may be understood as follows.

On one hand, there is a divergence between social and private costs that can lead to socially excessive suit. Specifically, when a plaintiff contemplates bringing suit, he bears only his own costs; he does not take into account the defendant's costs or the state's costs that his suit will engender.

On the other hand, there is a difference between the social and private benefits of suit that can either lead to a socially inadequate level of suit or reinforce the cost-related tendency toward excessive suit. Namely, the plaintiff does not recognize as a benefit to himself the social benefit of suit, its deterrent effect on the behavior of injurers. But the plaintiff does consider his private benefit, the gain he would obtain from prevailing. This private gain is not a social benefit but instead is a transfer from the defendant; it could be either larger or smaller than the social benefit. The contrast between the socially optimal and private incentive to sue was initially examined in Shavell (1982b)[105].

Let us consider the foregoing in more detail. Suppose that liability is strict. As stated, victims will sue if and only if

$$c_P < h.$$

Let x be the precaution expenditures that injurers will be induced to make if there is suit, q the probability of harm h if suit is not brought, and q' the probability of harm if suit is brought. (Thus, q' will be less than q if x is positive.) Suit will be socially worthwhile if and only if

$$q'(c_P + c_D + c_S) < (q - q')h - x,$$

where c_D is the defendant's litigation cost and c_S is the state's cost. In other words, suit is socially worthwhile if the expected litigation costs are less than the net deterrence benefits of suit. It is clear that the two foregoing conditions, for victims to sue and for suit to be socially optimal, are very different. Whether victims will sue does not

[105] See also subsequent examination of the issue in Menell (1983), Kaplow (1986b), Rose-Ackerman and Geistfeld (1987), and Shavell (1997).

depend on the costs c_D and c_S. Moreover, the private benefit of suit is h, the amount of harm (conditional on harm occurring), because this is what the victim will receive as a damages award; in contrast, the social benefit depends on the harm weighted by the reduction in the accident probability, $q - q'$, net of the cost of precautions x. It is evident, therefore, that victims might sue when suit is not socially optimal, and that victims might not sue even when suit would be socially optimal.

To illustrate the possibility of socially excessive suit, suppose that the losses a victim would suffer in an accident are \$10 000; that a victim's cost of bringing suit will be \$3000 and an injurer's cost of defending \$2000; that the probability of accidents is 10%, and that there is no precaution that injurers can take to lower the accident risk. Victims will then bring suits whenever accidents occur, for suing will cost a victim only \$3000 and yield him \$10 000. From the social perspective, this outcome is undesirable. Suit creates no beneficial deterrent, as injurers cannot do anything to lower risk. Yet suit does generate legal costs: expected legal costs are $10\% \times (\$3000 + \$2000) = \$500$. The bringing of suits is not socially desirable in this example because there are no incentives toward safety created by the suits. Yet this fact is of no moment to victims; nor are other parties' litigation costs. Victims bring suits for their private gain of \$10 000.

To illustrate the opposite possibility, suppose that the losses victims suffer in accidents are now \$1000, and an expenditure of only \$10 by injurers will reduce the probability of accidents from 10% to 1%. The costs of suit and of defending against suits are as in the previous example. In this case victims will not bring suits, as doing so will cost a victim \$3000 but yield him only \$1000. Hence injurers will have no reason to take care to reduce risk, and total expected social costs will be $10\% \times \$1000 = \$100$. It would be desirable for victims to bring suit, however. If they did, injurers would be led to spend \$10 to lower risk to 1%, and total expected social costs would be only $\$10 + 1\% \times (\$1000 + \$5000) = \$70$. Here the bringing of suits is socially worthwhile because of the significant reduction in the risk of accident losses that would result. (And observe that this is true even though the total legal costs of \$5000 exceed the victim's losses of \$1000.) But victims do not take the deterrence-related benefits of suit into account. Each victim looks only to his own gain from suit, which is small.

Under the negligence rule, the conclusions are qualitatively similar to those under strict liability, but the problem of excessive suit is less likely. To explain, assume initially that a victim would not sue a non-negligent injurer, because he would know that he would lose. Then it is socially desirable for victims always to be willing to bring suit against negligent injurers, however great the legal costs of suit would be. For if victims always stand ready to sue negligent injurers, injurers will be induced to act non-negligently. Thus, there will never actually be any suits for negligence – given the assumption that no one sues a non-negligent injurer – and thus no legal costs will be borne; deterrence of negligence will be achieved without legal cost. Although it is socially desirable for victims always to be willing to sue negligent injurers, victims of course will not do so if the cost of making claims exceeds their losses. Consequently, there might be a problem of too few suits.

Now assume, more realistically, that victims might sometimes bring suit against non-negligent injurers (or that injurers cannot perfectly control their behavior and sometimes act negligently). Then legal costs will in fact be incurred under the negligence rule. The situation will therefore be qualitatively similar to that under strict liability; there may be too many suits as well as too few, although one might suppose the problem of too many suits to be less severe than under strict liability.

It should be clear from our discussion that the point that the private and social incentives to bring suit may diverge is robust. On one hand, it will always be the case that the private cost of use of the system will be less than the social. And, on the other hand, the private benefits from suit will be what the plaintiff will win from suit, usually money, whereas the social benefits from suit will ordinarily be different: they will always include deterrence benefits and may also include compensation of victims (if insurance is unavailable) and the setting of precedent. These benefits litigants either will not take into account or will tend to weigh differently from their social importance[106].

5.1.3. Implications of the social and private divergence

The main implication of the social and private divergence is that state intervention may be desirable, either to correct a problem of excessive suit – notably, by taxing suit or barring it in some domain – or a problem of inadequate suit – by subsidizing suit in some way[107]. For the state to determine optimal policy, however, requires it to determine the effects of suit on injurer behavior and weigh them against the social costs of suit. It is thus not correct for the state to base policy on some simple, even though superficially appealing, criterion, notably, whether the plaintiff's expected gains from suit would have exceeded the aggregate litigation costs.

It should also be emphasized that the importance of the private–social divergence in incentives to sue may be substantial. This is suggested by the fact that the costs of use of the legal system are high; indeed, legal costs may on average actually equal the amounts received by those who sue[108]. Hence, the incentives created by the legal system must be significant to justify its use. However, regardless of whether the legal system creates valuable incentives, the private motive to bring suit may be great,

[106] Whereas this section was concerned with the implications of litigation costs for the frequency of suit, another litigation-cost related issue has to do with the level of precautions taken by injurers. The optimal level of precautions should reflect not only the direct harm that would be caused by an accident, but also litigation costs. To induce this higher level of precautions, injurers who are sued should pay not only the harm, but also (perhaps as a penalty) litigation costs borne by the plaintiff and by the state. On this issue, see Hylton (1990), Polinsky and Rubinfeld (1988a), and Shavell (1997, 1999).

[107] An additional response to the problem of inadequate suit that is sometimes possible is the use of class actions. See Dam (1975). Other approaches to encourage suits involve pro-plaintiff fee-shifting (see Section 5.4.2), damage multipliers [see Kaplow (1993)], and "decoupling" [see Polinsky and Che (1991) and Shavell (1997, 1999)].

[108] See Section 2.3.1.

giving rise to a reason for social intervention. Conversely, it may be important in some domains to create deterrence because this would have a significant effect on behavior, even though the money benefits of suit are too small for most victims to bring suit. This would justify the state's supporting litigation.

5.2. *Settlement versus trial*

Assuming that suit has been brought, we now take up the question whether parties will reach a settlement or go to trial[109]. A settlement is a legally enforceable agreement, usually involving a payment from the defendant to the plaintiff, in which the plaintiff agrees not to pursue his claim further. If the parties do not reach a settlement, we assume that they go to trial, that is, that some tribunal determines the outcome of their case. In fact, the vast majority of cases settle[110]. We discuss here two different models describing whether settlement occurs and then consider the socially optimal versus the private decision whether to settle.

5.2.1. *Exogenous beliefs model*

One model of settlement versus trial presumes that parties have each somehow come to a belief about the probability of the trial outcome; let p_P represent the probability of the plaintiff prevailing in his opinion, and let p_D be that same probability in the defendant's opinion. Let w be the amount that would be won (for simplicity assume that they agree about w). Assume also that the parties are risk neutral.

The plaintiff's expected gain from trial, net of litigation costs, is $p_P w - c_P$. This is the minimum amount he would accept as a settlement, rather than go to trial. The defendant's expected loss from trial, including his litigation costs, is $p_D w + c_D$; this is the maximum amount he would pay in settlement rather than go to trial. Hence, a settlement is possible if and only if $p_P w - c_P \leqslant p_D w + c_D$, in which case the settlement amount will be in the settlement range $[p_P w - c_P, p_D w + c_D]$. Note that if the parties agree on the probability p, the settlement range will be positive and $c_P + c_D$ in length. A settlement range does not exist, and trial will occur, when $p_P w - p_D w > c_P + c_D$. This means that the expected award in the plaintiff's opinion exceeds the expected award in the defendant's opinion by more than the sum of litigation costs. Thus, trial will tend to occur when the plaintiff is sufficiently more optimistic about winning than the defendant believes he should be[111].

[109] Cooter and Rubinfeld (1989), Daughety (2000), and Hay and Spier (1998) survey this general topic; Farmer and Pecorino (1996) review the asymmetric information literature on settlement versus trial.

[110] In fiscal year 1992, over 96% of civil cases in state courts did not go to trial; see Ostrom and Kauder (1996). In fiscal year 1995, almost 97% of federal civil cases were resolved without trial; see Administrative Office of the United States Courts (1995). These figures, however, overstate the settlement rate because many of the cases not tried were dismissed by a court rather than being settled. On the other hand, many disputes are settled before any complaint is filed.

[111] Loewenstein et al. (1993) and Mnookin (1993) suggest that litigant overoptimism is plausible.

Risk aversion of the parties will generally increase the size of the settlement range and thus, one presumes, makes settlement more likely. If the plaintiff is risk averse, he will be willing to settle for less than $p_P w - c_P$; and if the defendant is risk averse, he will be willing to pay more than $p_D w + c_D$.

The model under discussion originated with Friedman (1969), Gould (1973), Landes (1971), and R.A. Posner (1973), and was further elaborated in Shavell (1982c). It has the virtue of clarifying several basic intuitions: that settlement is fostered by litigation cost savings and by risk aversion, and that trial might result when plaintiffs expect to gain more than defendants expect to lose. The model also helps to explain the striking predominance of settlement in actuality. First, lawyers, who are experts on the law, are typically advising both litigants, and much information is acquired and comes to be shared by the opposing sides; we should thus expect beliefs of the two sides to be similar. Second, the costs of trial tend to be substantial. These observations suggest that a settlement range typically exists and thus that settlement would be likely to occur[112].

The model has the additional virtue of being simple and easy to manipulate because it focuses on the calculation of the settlement range[113]. The model is unsatisfying, however, in two respects. It does not explain the origin of parties' beliefs. And it does not include a description of rational bargaining between the parties; thus, it does not explain whether there will be a settlement when there is a positive settlement range or the amount of any settlement within the range.

5.2.2. Asymmetric information model

A second type of model of settlement versus litigation presumes that there is asymmetry of information between litigants, and includes an explicit account of bargaining. The simplest of such models is that of Bebchuk (1984), in which there is one-sided asymmetry of information and bargaining consists of a single take-it-or-leave-it settlement offer made by the party without private information[114]. Suppose,

[112] The observations also raise interesting questions about the timing of settlement – will it occur early or late? (In fact, many cases settle early, but many also settle late, on the eve of trial or even during trial.) A reason for settlement to be delayed is that at the outset of settlement negotiations, information may be disparate; but, as noted, over time, as information is acquired and shared, the parties' beliefs tend to converge. A reason for settlement to occur early, however, is that this maximizes the parties' savings in litigation costs. To express the point differently, as time passes, more litigation costs are sunk, meaning that the savings from settlement are lowered, tending to decrease the chances of settlement. For analysis of the timing of settlement (in a model of asymmetric information), see Spier (1992a).

[113] For example, the larger the possible judgment amount w, the greater the chance of trial, for a larger judgment magnifies the effect of differences of opinion in the likelihood of trial outcomes ($p_P w - p_D w$, when positive, is increasing in w). However, larger judgments tend to reduce the likelihood of trial if litigants are risk averse.

[114] Asymmetric information models of trial versus settlement have been refined and extended in various ways. See, for example, Daughety and Reinganum (1994) (in which asymmetry of information is two-

for example, that the defendant has private information about the probability p that the plaintiff will win at trial (perhaps the defendant possesses private information bearing on whether he will be found negligent)[115]. The plaintiff makes a settlement offer x, knowing that low-p defendants will reject his offer and high-p defendants will accept; specifically, if $pw + c_D < x$, the defendant will reject and the plaintiff will therefore obtain only $pw - c_P$, but if $pw + c_D \geq x$, the defendant will accept and pay x. The plaintiff, who knows the probability distribution over p, chooses x to maximize his expected payoff from settlement or trial[116]. The higher his offer x, the more he will obtain if his offer is accepted, but the greater is the likelihood of rejection and thus of his bearing trial costs. At the optimal offer for the plaintiff, there will generally be a positive probability of trial and also of settlement. Furthermore, it can be shown that the higher are litigation costs, the more likely is settlement, and that risk aversion increases the likelihood of settlement.

This model, note, is roughly consistent with the previous one of Section 5.2.1 in the sense that trial occurs due to disparate beliefs (arising out of the asymmetry of information). In particular, the plaintiff's opinion of the probability of winning is the mean probability $E(p)$ over the distribution of defendants, and trial will occur if the defendant's p is sufficiently low in the distribution. In addition, the comparative statics of the present model are similar to that of the previous one (for instance, as just noted, higher litigation costs make settlement more likely).

The primary virtues of such asymmetric information models are twofold. First, they include an explicit account of bargaining and thus of the probability of settlement and the magnitude of the settlement offer. (But the ability to predict the probability of settlement and the magnitude of the settlement offer is to some extent specious. Under the bargaining models studied, essentially arbitrary modeling choices are made over such matters as who makes the offer, the informed or the uninformed party; these choices substantially influence the probability of settlement and the settlement offers[117].) Second, the models explain differences of opinion that give rise to trial in terms of differences in possession of information. (However, the models do not explain

sided), Hay (1995) (in which unobservable case preparation contributes to asymmetry of information), Reinganum and Wilde (1986) (in which the informed party makes the offer, and the uninformed party makes an inference from it), Schweizer (1989) (in which asymmetry of information is two-sided), and Spier (1992a) (in which there are multiple rounds of bargaining and, as discussed in note 112, the focus is on the timing of settlement). For a useful survey of asymmetric information models of litigation, see Farmer and Pecorino (1996), and for a general survey of asymmetric information models of bargaining, see Kennan and Wilson (1993). For empirical investigations of litigation that emphasize asymmetric information, see Farber and White (1991), Osborne (1999), and Sieg (2000).

[115] Asymmetric information could also concern the magnitude of the judgment or factors independent of the trial itself, such as parties' degree of risk aversion, their short-run need for funds, their tastes for litigation, and, as mentioned in the previous note, unobservable aspects of case preparation.

[116] Specifically, the plaintiff's expected payoff as a function of x is $\int_0^z (pw - c_P) f(p)\,dp + (1 - F(z))x$, where $z = (x - c_D)/w$, f is the density of p, and F is the cumulative distribution of p.

[117] For an attempt to address this problem, see Daughety and Reinganum (1993).

why there should be such differences in information, especially given the incentives for sharing of information and the possibility of its forced disclosure through legal discovery; we discuss these issues below in Sections 5.4.7 and 5.4.8.)

5.2.3. Socially optimal versus privately determined settlement

The private and the social incentive to settle may diverge for reasons related to those explaining the difference between the private and the social incentive to sue (see Section 5.1.2)[118]. First, because the parties involved in litigation do not bear all the costs of a trial – the salaries of judges and ancillary personnel, the forgone value of juror time, implicit rent on court buildings – the parties save less by settling than society does, which suggests that the private incentive to settle is socially inadequate. Second, when there is asymmetric information, parties will fail to settle – and thus litigation costs will be incurred – when their demands turn out to have been too aggressive. But their desire to obtain from each other a greater share of their litigation cost savings does not itself translate into any social benefit. Third, the prospect of settlement may reduce deterrence because defendants gain from settlement. This need not, however, be socially undesirable because settlement lowers the real total social cost of harmful acts, making less deterrence appropriate[119]. Also, the division of surplus in settlement may affect deterrence. Fourth, the prospect of settlement may increase deterrence because it lowers plaintiffs' expected litigation costs and thus increases the chance of suit. These latter two factors are not, of course, taken into account by the parties to settlement negotiations. Finally, by averting trial, settlement may have other effects on social welfare. For example, trials may reveal socially valuable information (such as about product hazards that consumers could guard against) or lead to new precedents. These are also factors that parties may ignore or treat inappropriately (a firm might have a socially perverse incentive to avoid trial to conceal information about product hazards).

The state can act to correct a divergence between private and social incentives to settle. A point that should be stressed in considering optimal social policy is that if settlements were to reduce deterrence undesirably, this does not imply that trial should be fostered; deterrence could be enhanced by raising damages to induce settlements for greater amounts or by imposing a tax on defendants (regardless of whether they settle). Trial is desirable only when there is no less costly way to raise social welfare, and a conjecture is that the usual social problem is that there are too many trials, not too few[120].

[118] The normative question concerning the social versus the private value of settlement has received little attention relative to the positive question of when parties will settle. On the normative question, see Polinsky and Rubinfeld (1988b), Shavell (1997, 1999), and Spier (1997).

[119] Settlement reduces ex post social costs by the sum of the plaintiff's, defendant's, and court's costs, but deterrence is reduced in a way that reflects only the fraction of the savings in the plaintiff's and defendant's costs that is captured by the defendant in settlement bargaining.

[120] In fact, courts attempt to promote settlement in a variety of ways.

5.3. Litigation expenditures

5.3.1. Private incentives to spend on litigation

Here we focus on litigant expenditures given that suit has been brought. (We should note that litigation expenditures are made prior to trial as well as during trial; indeed, most are incurred in cases that settle.) Suppose that each litigant's expenditures are made noncooperatively, as in Braeutigam, Owen and Panzar (1984), Katz (1987, 1988), and R.A. Posner (1973). Under this assumption, a plaintiff will make litigation expenditures as long as this raises his expected return from settlement or trial (net of litigation costs), and a defendant will make such expenditures as long as this lowers his expected total outlays. The effects of each litigant's expenditures will generally depend on what the other does; indeed, the two will often be spending to rebut one another[121].

5.3.2. Social versus private incentives to make litigation expenditures

There are several sources of divergence between social and private incentives to spend during litigation. First, as just noted, the litigants may well be spending in ways that offset each other. To the extent that their expenditures do not alter trial or settlement outcomes, the expenditures constitute a social waste. Second, even if expenditures are not offsetting, they may mislead the tribunal rather than enhance the accuracy of outcomes. Such expenditures have negative social value.

Third, even if expenditures do improve the accuracy of outcomes, they may not be socially optimal in magnitude. By analogy to what we stressed in Section 5.1.2, the parties decide on their expenditures based on how they influence the litigation outcome, without regard to their influence (if any) on incentives. This could lead to expenditures that are too great or too small, relative to what is socially correct[122].

An important instance of the possibility that expenditures could be socially excessive concerns the assessment of damages. See Kaplow and Shavell (1996b). Suppose that the presently estimated harm deviates from the truth by $100. Then one of the litigants will be willing to spend up to $100 to prove the correct amount (it will be the defendant if the estimate exceeds the correct level, and the plaintiff if the estimate falls below the correct level). It can be shown that the social value of the more accurate estimate tends, however, to be lower than $100, because the social value of accuracy is based on its effects on incentives. Indeed, there will sometimes be no beneficial incentive effect from more accurate assessment of harm, such as when errors (in the absence of

[121] One may contrast a system in which a single authority, perhaps the tribunal, makes decisions about litigation expenditures. This is done to an extent in many European countries for criminal proceedings. In the United States, federal trial court judges occasionally use special masters or court-appointed experts to perform similar functions.

[122] See generally Kaplow (1994a).

additional expenditures) are unbiased and not predictable ex ante by potential injurers. In particular, potential injurers, at the time they choose their precautions, will often know only a probability distribution of possible harm, so litigation expenditures ex post that provide a precise assessment of a particular victim's actual harm would not affect incentives[123].

Expenditures on determining whether a party is liable (as opposed to the magnitude of damages) could be socially excessive or inadequate[124]. To illustrate the latter possibility, suppose that the cost of establishing that a defendant was negligent exceeds the amount of harm suffered. Plaintiffs will not have an incentive to make the necessary expenditure, with the result that negligence might not be discouraged. But if the deterrent effect of liability were significant, that result would be undesirable. (Suppose that deterrence would eliminate most negligently caused harm, so that ex post litigation costs would not often have to be incurred; see Section 5.1.2.)

Because private and social incentives to spend on litigation may diverge, it may be beneficial for expenditures to be either controlled or encouraged. In practice, courts often act to restrict the legal effort that parties can undertake, for example, by limiting the extent of discovery and the number of testifying experts.

5.4. Extensions of the basic theory

We consider here various extensions of the basic theory discussed above; for the most part, these extensions are concerned with the description of litigation rather than with its normative analysis[125].

5.4.1. Nuisance suits

A nuisance suit is often defined as a suit that the plaintiff brings even though he would not actually pursue his case to trial, because the expected award he would obtain is less than the trial cost; in this sense a nuisance suit is a negative expected value suit. We should first point out that we cannot infer that nuisance suits should not be brought: as we stressed in Section 5.1.2, it is possible that the social deterrence benefits of a type of suit make it desirable to bring even though litigation costs exceed the expected

[123] See Section 2.4.4. If compensation of risk-averse victims were important due to the unavailability of insurance (see Section 2.2), more accurate compensation will have social value, although parties' incentives to make litigation expenditures would still tend to be excessive because the insurance benefit of avoiding a $1 error in compensation is less than the maximum of $1 that a party would be willing to expend to correct the error. See Kaplow (1994b).

[124] The determination of liability in the context of law enforcement (see Section 6) is analyzed in Kaplow and Shavell (1994a).

[125] For empirical studies that bear on the theory of litigation, see, for example, Hughes and Snyder (1995), Ramseyer and Nakazato (1989), and Viscusi (1986a, 1988).

judgment[126]. (Nor, as we emphasized, can it be assumed that non-nuisance suits – positive expected value suits – ought to be brought.)

A major question about nuisance suits, and the one to which primary attention has been given, is why they are brought in view of their negative expected value. One important explanation concerns asymmetric information: that plaintiffs who are not willing to go to trial are not identifiable to defendants and ride on the coattails of plaintiffs who would be willing to go to trial. As a consequence, the plaintiffs who are unwilling to go to trial are able to settle for a positive amount with defendants; see Bebchuk (1988) and Katz (1990a). Another possibility, not premised on asymmetric information, is that a plaintiff can initiate a suit at low cost and, although he would lose if the defendant undertook substantial litigation effort, he would prevail if the defendant did not. In this case, the defendant might prefer to settle to avoid paying defense costs; see Rosenberg and Shavell (1985). An additional reason concerns the point that, as plaintiffs spend continuously on litigation, their willingness to go to trial increases (because the amount that they would then save by not going forward diminishes); see Bebchuk (1996)[127].

5.4.2. Shifting of legal fees

Thus far, we have assumed that parties bear their own legal costs, a regime referred to as the *American rule.* By contrast, under the *English rule,* the loser pays the legal costs of both sides. Fee-shifting may also be one-way, favoring the plaintiff (that is, shifted only to the defendant, if the plaintiff wins) or favoring the defendant (shifted only to the plaintiff, if the defendant wins)[128]. Fee-shifting has clear implications for the incentive to sue; for example, under the English rule, suit is encouraged, relative to a regime of no fee-shifting, if the plaintiff's probability of winning is sufficiently high, because then his expected costs of trial fall[129].

Fee-shifting may increase the chance of trial given that suit has been brought, essentially because it accentuates differences in litigant estimates of the expected gains and losses from trial; see R.A. Posner (1977a) and Shavell (1982c). Under the English rule (the effects under one-way fee-shifting are similar), if the plaintiff and the defendant are each optimistic about winning, then each will be optimistic about

[126] This is obviously not to deny that some types of nuisance suits are undesirable. For example, where plaintiffs would not prevail because their cases are fictitious, their bringing of suits would tend to distort incentives as well as waste resources on litigation.

[127] There is also a literature on how nuisance suits (or, relatedly, suits with a low probability of success) might be discouraged. See Bebchuk and Chang (1996), Katz (1990a), and Polinsky and Rubinfeld (1993, 1996).

[128] For a description of the use of fee-shifting, see Derfner and Wolf (1995). Fee-shifting favoring only plaintiffs is used to stimulate suits where the private incentive is thought to be inadequate, whereas fee-shifting favoring only defendants is usually proposed as a means to discourage frivolous litigation.

[129] Fee-shifting may also affect injurers' incentives to reduce risk and thus the number of harmful outcomes that could lead to suits. See van Wijck and van Velthoven (2000).

passing on his legal expenses to the other, which tends to reduce the settlement range and increase the chances of trial[130]. However, fee-shifting tends to raise the amounts the parties will spend at trial, as a party's expenditure will only be a cost to him with a probability rather than with certainty[131]. This increase in anticipated litigation expenditures attenuates the rise in the chance of trial. Also, fee-shifting makes trial riskier, so that if parties are risk averse, it may reduce the chance of trial. Because of the variousness of the effects of fee-shifting (and because of the divergences between private and social incentives to sue, to settle, and to spend on litigation), its influence on social welfare is generally ambiguous, as is emphasized, for example, in Gravelle (1993).

A variant of simple fee-shifting is an offer-of-settlement scheme, according to which fees are shifted only if a settlement proposal is rejected and the amount actually awarded differs in a specified way from the rejected proposal. For instance, if a defendant rejects a plaintiff's offer and the actual trial award exceeds that offer, fees might be shifted to the defendant. The effects of such schemes on settlement are complex and not readily summarized. See Bebchuk and Chang (1999), Miller (1986), and Spier (1994a).

5.4.3. Additional elements of trial outcomes

We have assumed that the only outcome of a trial is a judgment paid by the defendant and received by the plaintiff, but there are other possibilities. First, a trial outcome may have implications for a litigant beyond the immediate judgment. For example, a firm may believe that a loss at trial would invite a string of future lawsuits; thus, a loss would be more costly for it than the judgment[132]. This would tend to make settlement more likely, as it would raise the amount the defendant firm would be willing to pay in settlement. Second, a litigant may care whether a trial is held per se: a plaintiff might, say, wish the defendant to be exposed to public scrutiny. This would make trial more likely. Or a party might want to avoid a trial because it would result in the airing of embarrassing facts or the disclosure of valuable business information, which would tend to make trial less likely. Third, in cases such as child-custody disputes, the combination of indivisibilities and wealth constraints may make settlement less likely[133].

[130] The same conclusion holds, for closely related reasons, in the asymmetric information model of Bebchuk (1984).

[131] See Braeutigam, Owen and Panzar (1984), Hause (1989), and Katz (1987). In Katz's simulation, the English rule increases costs by 125 percent.

[132] See, for example, Che and Yi (1993). A party's willingness to settle and the amount of settlement may also have reputational effects. See Miceli (1993).

[133] See Shavell (1993b). For instance, suppose that for each parent in a custody dispute, the value of custody is equivalent to $1 000 000, each parent believes custody would be awarded with probability 50%, and the cost of trial for each is $10 000. Then to induce either parent to settle and give up the opportunity

5.4.4. Statistical inference from cases that go to trial

A question of interest is whether cases that go to trial are representative of the underlying population of cases, and notably, whether the likelihood of plaintiff victory at trial or the amounts won are typical of the cases that settled. This question is important because, often, the most readily available data is on cases that go to trial, whereas the great majority of cases settle. As Priest and Klein (1984) first emphasized, the cases that go to trial may be quite different from settled cases. For example, if in 99% of cases defendants would be found liable for a certain amount, but in 1% of cases defendants would prevail, then, if plaintiffs cannot distinguish the two groups, plaintiffs will likely insist on a settlement amount that the former defendants would pay and the latter would reject. Hence, defendants would win all cases that go to trial, which would be wholly unrepresentative of the cases that settled. In general, cases that go to trial are not representative of the underlying population of cases, and the proper manner of making inferences from trial outcomes is complex [134].

5.4.5. Lawyers as agents of litigants

Because clients and their lawyers are in a principal and agent relationship, the general problems of principals and agents are relevant for clients and lawyers. Consequently, to the degree that clients cannot observe lawyers' effort levels and lack legal expertise, a fee arrangement linked to lawyers' performance might have joint value to them, but it would impose risk on lawyers (although it would simultaneously reduce clients' risk, and many clients – particularly individuals or small entities – may be risk averse) [135].

In fact, lawyers often are compensated at an hourly rate for time spent, without regard to legal outcomes. The only important explicit exception is that plaintiffs' lawyers in tort actions frequently are paid a fraction of the amount they obtain for their clients under a so-called contingent fee agreement [136]. In addition, lawyers are

of custody, an offer of at least $490 000 would have to be made, yet neither parent may have assets nearly equal to that amount. Thus, despite the fact that the parents agree about the likelihood of trial outcomes and could save litigation costs by settling, they would go to trial.

[134] Priest and Klein (1984) suggested that cases that go to trial would be won by plaintiffs approximately 50% of the time, regardless of the underlying population of cases. This somewhat surprising conclusion of theirs is correct given their assumptions; but it is not borne out in fact, and does not hold under general assumptions about the population of cases and bargaining over settlement and trial. See Eisenberg (1990), Eisenberg and Farber (1997), Hylton (1993), Shavell (1996), Waldfogel (1995b), and Wittman (1985).

[135] Another problem in the agency relationship is that, at the time of contracting, the client may not know the lawyer's quality, and there may also be asymmetric information regarding the strength of the client's case. For a discussion of how these problems may affect fee arrangements, see Dana and Spier (1993), Emons (2000), and Rubinfeld and Scotchmer (1993).

[136] See generally Rubinfeld and Scotchmer (1998) on contingent fees. We note that payment arrangements that are contingent upon outcomes may be common because lawyers who nominally charge hourly rates may submit higher bills when successful and may trim their bills when they lose.

implicitly rewarded on the basis of performance in the sense that they (and their firms) acquire reputations, so that their future business depends on performance. Lawyers' conduct is also controlled to some extent by the threat of suit by clients for malpractice, by court-mandated penalties, and by bar association discipline. See Wilkins (1992).

Principal–agent problems that are specific to the legal context arise in the decisions to sue and to settle versus go to trial. See Miller (1987). For example, when lawyers are paid on a contingent fee basis, they might have perverse incentives to favor not bringing suits or to settle, because their own gain would be only a fraction of the total gain from winning. See Danzon (1983) and Hay (1996)[137]. When lawyers are paid on an hourly basis, it is often said that they have an excessive incentive to sue and to reject settlement offers in favor of trial. (This claim, however, assumes that their hourly rate exceeds their opportunity costs; if, for example, additional, more profitable work comes into the office after the hourly rate is set, then hourly-compensated lawyers may have an excessive incentive to settle.)

5.4.6. Insurers as agents of litigants

Insurers often play a role in litigation. In accident suits, for example, plaintiffs may own medical or disability insurance policies with clauses giving their insurers the right to bring suit and conduct litigation, and defendants frequently hold liability insurance policies that give insurers a role in litigation. Conflicts may arise between litigants and their insurers as their agents in litigation when the coverage ceiling is less than the amount at stake in litigation. See Meurer (1992) and Sykes (1994). To illustrate, suppose that there is a 20% chance that trial would result in a finding of liability and that losses are \$500 000; assume also that the defendant's liability coverage ceiling is \$150 000. The liability insurer would prefer to reject a settlement offer of \$75 000, even though the offer falls below the expected judgment of \$100 000. (If the settlement offer is accepted, the insurer pays \$75 000 for sure, whereas if there is a trial the insurer makes a payment of \$150 000 only 20% of the time, which has an expected cost of \$30 000.) By similar reasoning, a plaintiff's insurer would tend to want to settle for less than plaintiffs would like, which would increase the chance of settlement. Note, however, that reputational interests of insurers, as well as the possibility of renegotiation between insurers and insureds, serve to mitigate their conflicts of interest.

5.4.7. Voluntary sharing of information

In the discussion of settlement versus trial in Section 5.2, we assumed that the information of parties was somehow exogenously determined: either information was

[137] Hay (1997) discusses how bifurcated contingent fees (which pay a higher rate if there is no settlement) can help to address this problem. Additional principal–agent problems arise in the context of class actions, where many plaintiffs are joined in a class and there is a free-rider difficulty with regard to supervision of attorneys. See Coffee (1986) and Macey and Miller (1991).

in the background and formed parties' perhaps disparate beliefs, or else information was explicitly presumed to be asymmetric. However, litigants in general have strong motives to share information. See Shavell (1989a). Most obviously, parties will want to share favorable information in order to foster settlement and to improve its terms. A plaintiff, for example, would want to show the defendant information establishing that his losses were in fact higher than the defendant otherwise believes; in this way, the plaintiff can induce the defendant to pay more in settlement and perhaps avoid an impasse leading to trial. Likewise, a defendant would want to show the plaintiff evidence pointing toward his lack of responsibility, in order to convince the plaintiff to accept a lower settlement offer.

In addition, parties will want to reveal information to avoid negative inferences that would be made from their silence. If a plaintiff says nothing about the magnitude of his losses, the defendant will be likely to infer that the plaintiff is withholding information that his losses are lower than average, and if this inference is made, the defendant will not be willing to make an average offer. Both this incentive to avoid negative inferences and the incentive to reveal favorable information tend to produce significant voluntary disclosure and help to explain the high rate of settlement[138].

Nevertheless, certain information will not be shared, and this helps to explain why some cases do not settle. First, a party may decide against disclosing information because revealed information can often be countered at trial if the opposing side has foreknowledge of it. Second, information may be difficult to share, even though a party wants to do that. For instance, a plaintiff might know that his business losses from a breach of contract will be high, but not be able to demonstrate this during settlement negotiations (because, say, experts will have to be hired for trial to verify the losses). Another difficulty faced by a party who wants to reveal favorable information is that it may consist of the absence of unfavorable information. (For example, if the defendant was not drinking before a traffic accident, his favorable information may be the nonexistence of anyone who saw him drinking, and he may have no way to demonstrate this[139].) Third, information may not be shared because it is unfavorable and the negative inference drawn from silence is not too strong. Note that the negative inference from silence will be weakened to the extent that some parties do not disclose favorable information for the first two reasons just given.

5.4.8. Required disclosure of information – legal discovery

The courts may require that a litigant disclose certain information to the other side; this practice is known as discovery. It is commonly believed that discovery

[138] Farber and White (1991) find that many malpractice cases settle after plaintiffs obtain information from defendants.

[139] If the case does not settle, the plaintiff may ultimately be able to verify the defendant's claim implicitly: investigations may fail to locate any person who saw the defendant drinking (whereas if there really is a witness, there is some probability that the witness would be located).

significantly increases the likelihood of settlement because it reduces differences in parties' information. But, as just emphasized, there may well be substantial voluntary sharing of information, so the influence of compulsory disclosure will not be so great and is in fact nonexistent in a natural model of disclosure. See generally Shavell (1989a), and see also Hay (1994)[140].

Discovery will, nevertheless, tend to increase the rate of settlement and also will affect the terms of settlements. First, when parties would otherwise withhold favorable information to disable the opponent from countering it at trial, discovery will force disclosure, which in turn will make settlement more likely. Second, when parties would otherwise withhold unfavorable information (because the negative inference from so doing would not be too strong), discovery will mandate disclosure and lead to settlement on less favorable terms. It should be noted, however, that such parties with unfavorable information would have settled in the absence of discovery[141]. Settlement will increase overall because, when those with unfavorable information are required to disclose it, more generous offers will be made to those who remain silent in the face of discovery (perhaps those with favorable information who cannot verify the strength of their cases). Third, the prospect of legal sanctions for false statements may make more credible parties' insistence that they lack certain unfavorable information (such as the assertion that there is no witness who could testify to the party having been drinking before an accident); this would encourage settlement of such cases[142].

Discovery may also be used strategically. Obeying discovery requests is often expensive because significant time and resources may be needed to produce the desired information. This fact raises questions about the use of discovery requests as a threat, for the costs of compliance with discovery requests are, under our current system, generally borne by the side asked to comply. It also raises questions about the socially optimal amount of discovery.

[140] Other models of discovery are Sobel (1989) and Mnookin and Wilson (1998); because these articles do not compare outcomes when there is discovery with outcomes with voluntary sharing of information, they are hard to interpret. See also Cooter and Rubinfeld (1994), a model of discovery without an explicit treatment of asymmetric information, Schrag (1999), a model in which there is judicial control of discovery, and Shepherd (1999), an empirical study of discovery.

[141] Those who withhold unfavorable information when there is no compulsory discovery seek to mimic others with favorable information who remain silent (whether because they strategically withhold information or because they cannot credibly verify their favorable information). Accordingly, they receive settlement offers that reflect the average characteristics of the silent group; being those in the group with the least favorable cases, they will be the ones who settle, on terms that are better than they can expect if they were to disclose their unfavorable information. See Shavell (1989a).

[142] There may, however, be limitations on the feasibility of enforcement of discovery obligations. If a side fails to divulge unfavorable information, often this will not come to light (because the case may settle beforehand or because, even if there is a trial, the other side may never learn the truth in any event). Accordingly, very high sanctions for misrepresentations and possibly selective investigation (perhaps by the state) of the veracity of discovery responses may be necessary, although the present system does not follow either course.

5.4.9. Criminal adjudication

The analysis of suit and settlement for criminal adjudication [see, for example, Landes (1971) and the literature cited in Section 6.3.8 on plea bargaining] is in some respects similar to that for civil adjudication, but there are differences in parties' incentives that are worthy of note. First, in criminal cases the complaining party is a public prosecutor. Accordingly, litigation decisions will not be based on a simple comparison of litigation costs and the expected gain because the prosecutor neither directly bears these costs nor benefits monetarily from winning (costs are borne by the state and there is no actual recovery). Instead, a prosecutor's decisions will be dictated by the complex of factors determining his salary and his professional future. Nevertheless, one expects there to be a rough congruence between prosecutorial behavior and what the basic theory suggests. For example, prosecutors should tend to bring cases that have higher prospects of success and are less costly, and asymmetric information may impede settlement.

Second, a criminal defendant is often impecunious and will have been assigned a public defender. Not having to pay for his defense, such a defendant will not save legal expenses by settling, making him less willing to settle than otherwise. But those who serve as public defenders or are appointed to represent indigent defendants will often have limited budgets and receive low compensation, so they may exert less effort than what defendants would demand were they not liquidity constrained. Also, criminal defendants, and especially first-time defendants, may not care so much about the magnitude of punishment as about the fact of a criminal conviction or about having to spend any time in prison. If so, they would be less willing to settle than otherwise. There are other possibilities, of course, but our main point is that the basic theory provides only a very rough prediction of suit (prosecution) and settlement in the criminal context.

5.4.10. Additional aspects of legal procedure

There are many aspects of legal procedure that merit study but which we do not examine here, due mainly to their having received only limited treatment in the literature. Topics include the burden of proof[143], rules of evidence (and tribunals' making inferences from evidence)[144], the use of juries[145], the behavior of judges[146], summary adjudication[147], class actions[148], sequential versus joint

[143] See Davis (1994), Hay and Spier (1997), Kaplow (1994a), R.A. Posner (1973), Rubinfeld and Sappington (1987), and Sobel (1985).

[144] See Daughety and Reinganum (1995), Froeb and Kobayashi (1996), Lewis and Poitevin (1997), R.A. Posner (1999), Schrag and Scotchmer (1994), and Shavell (1989b).

[145] See Klevorick, Rothschild and Winship (1984) (jury deliberations), and Schwartz and Schwartz (1996) (peremptory challenges).

[146] See Cohen (1992), Elder (1987), Higgins and Rubin (1980), Kimenyi et al. (1993), Kornhauser (1992a, 1992b), R.A. Posner (1993a), Ramseyer (1998), and Rasmusen (1994).

[147] See R.A. Posner (1986) on summary jury trials.

[148] See Che (1996), Dam (1975), Miller (1998), and Silver (2000).

adjudication of multiple issues in a single case[149], the sharing of liability among multiple defendants[150], the use of lawyers as advocates for clients[151], and the adversarial system of adjudication (in which each side substantially controls its litigation activity) versus the European inquisitorial model (in which the tribunal controls much litigation activity)[152].

5.5. Legal advice

Because legal advice is costly, individuals must make decisions whether or not to obtain it, and questions about its social desirability also arise. In discussing the topic of legal advice, it is useful to consider separately ex ante legal advice – obtained when a party is contemplating an action – and ex post legal advice – secured after a party has acted or someone has been harmed, which is to say, at the stage of possible or actual litigation. A notable difference between the types of advice is that ex ante advice can channel behavior directly in conformity with law, whereas ex post advice comes too late to accomplish that (although it has indirect effects on behavior). Ex ante legal advice was first studied from an economic perspective in Shavell (1988) and Kaplow and Shavell (1992); ex post legal advice was initially investigated from this standpoint in Kaplow and Shavell (1989, 1990)[153].

5.5.1. Ex ante legal advice: when acts are contemplated

Advice has private value to a party who is considering taking some action with a possible legal consequence if the advice might lead him to alter his decision. The private value of legal advice is just an instance of the conventional definition of the expected value of information to a decisionmaker, as presented for instance in Raiffa (1968).

The social, as opposed to the private, value of ex ante legal advice inheres in the social desirability of advice-induced changes in parties' behavior. In general, advice has positive social value because it promotes adherence to legal rules. The specific nature of the comparison between the social and the private values of legal advice depends on the form of liability. When liability is strict, the private value of legal advice is the same as its social value. This basic conclusion follows essentially because a party's liability burden equals the harm he causes. When, however, liability is based on negligence, the private value of legal advice can be shown to exceed its

[149] See Landes (1993).

[150] See Easterbrook et al. (1980), Klerman (1996), Kornhauser and Revesz (1994), and Polinsky and Shavell (1981).

[151] See Dewatripont and Tirole (1999).

[152] See, for example, Daughety and Reinganum (2000a), Langbein (1985), and Shin (1998), and, closely related, Milgrom and Roberts (1986).

[153] Legal advice was further studied in Bundy and Elhauge (1991, 1993) and Fischel (1998).

social value. The explanation is in part that if a person avoids negligence because of advice, his liability saving will generally be larger than the reduction in expected harm he accomplishes, for he will escape liability entirely even though his non-negligent behavior might still cause harm.

5.5.2. Ex post legal advice: at the stage of litigation

The private value of ex post legal advice resides in the possibility that the advice will lead a party to change his decisions about suit, settlement, and trial.

In considering the social value of ex post legal advice, observe first that because such advice is, by its nature, imparted to parties only after they have acted, it cannot have aided them initially in conforming with the law. A firm that does not know whether discharging a chemical waste into a river will violate an antipollution statute obviously cannot be led to behave appropriately by learning what the law is after it decides about discharging the chemical. This simple but fundamental observation means that ex post advice does not raise social welfare in the direct way that ex ante advice does. Nonetheless, ex post advice certainly may influence behavior and social welfare.

Ex post advice that defendants obtain in the course of a lawsuit may affect social welfare by lowering sanctions for those who knowingly violate the law, that is, ex post advice may dilute deterrence of undesirable conduct. Lawyers may lower expected sanctions by advantageous use of legal strategy and, importantly, by counseling defendants on the selection of evidence to present and to suppress. Given that individuals anticipate that their expected sanctions for causing harm will be reduced due to the subsequent availability of legal advice, fewer individuals will be deterred from engaging in undesirable behavior. Thus, legal advice may have negative social value, a point that was early emphasized by Bentham (1827). This reasoning, however, is incomplete, in part because the state may be able to raise overall sanctions to offset the dilution of deterrence due to advice.

Ex post advice may, however, enhance social welfare by increasing otherwise inadequate sanctions that would be imposed on those who knowingly commit sanctionable acts. Specifically, advice may raise expected sanctions because lawyers may help plaintiffs to obtain higher judgments, better reflecting the harms they have sustained. Additionally, ex post advice may raise social welfare by lowering sanctions for defendants who did not violate the law, or who face higher sanctions than they should.

There is thus no way on the basis of logic alone to conclude whether or not ex post advice provided during litigation is on balance socially desirable – whether or not its socially undesirable effect, due to dilution of deterrence, is less important than its desirable effect, due to increased accuracy of legal outcomes for the guilty and for the innocent. Moreover, it is not obvious whether the net effect of advice will be to increase or to decrease the accuracy of adjudication.

Let us, however, restrict attention to ex post legal advice that does increase the accuracy of legal outcomes, and ask how the typically positive social value of this

advice compares to its private value. The general answer to this question is that either the private value of the advice or its social value could be larger, so that the private incentive to spend on the advice could be socially excessive or it could be inadequate. The reason is essentially that explained in Section 5.1.2; the social value of legal advice that increases accuracy inheres in its incentive effect on prior behavior of parties, and this has little connection to the private incentive to spend on advice, for that derives from the amount at stake in litigation. In some contexts, however, the private value of accuracy-enhancing advice will tend to exceed the social value and too much will be purchased. Notably this may often be true of advice about proving the extent of harm, for the reasons we explain in Section 5.3.2.

In sum, the social value of ex post legal advice is complicated to determine, possibly negative and possibly positive, and not closely related to its private value. In certain domains, a plausible conjecture is that, in an appropriate average sense, the private value of ex post advice exceeds its social value.

5.5.3. Other aspects of legal advice

Subversion of the law. One issue that we have not mentioned is that advice may directly subvert the law. Lawyers may lower the effective magnitude of sanctions by helping clients to hide assets, and lawyers may also decrease the likelihood of sanctions if they have knowledge of enforcement strategies (such as how the tax authorities choose whom to audit). Of course, lawyers are not supposed to thwart law enforcement, but they have an economic incentive to do so and can fairly easily avoid punishment for it (lawyers give advice in private and can phrase their advice in hypothetical but readily understood terms). From the social perspective, legal advice that frustrates law enforcement is obviously undesirable.

Confidentiality of legal advice. The legal system protects the confidentiality of communications between lawyers and their clients under wide circumstances. Confidentiality of legal advice will benefit clients when there is a positive probability that disclosure of advice would lower its value to them. This would usually be true of advice about the selection of evidence to present in litigation: such advice generally would be robbed of effectiveness if it were disclosed to the opposing side and the court. Confidentiality is also of obvious importance to those obtaining advice subversive of the law. By contrast, confidentiality often should not matter to parties obtaining advice about the legality of an act or about the magnitude or the likelihood of sanctions, because disclosure of such advice will usually not disadvantage them. (This is because they seek advice with the intention of following the law[154].) Still, whatever is the character of strictly legal advice, maintaining the confidentiality of much business or personal information about clients themselves will frequently be of importance to the clients.

[154] See Shavell (1988).

Because protection of confidentiality can benefit clients (and never is a disadvantage to them), it encourages clients to consult with and reveal information to their lawyers. This in itself is sometimes thought to imply that confidentiality is socially desirable. That reasoning, however, is mistaken: confidentiality is socially desirable only if the legal advice that confidentiality encourages is socially desirable, and, as has been explained above, that may not be the case.

Confidentiality of legal work product. The legal system also protects the confidentiality of legal work product (documents and other records of lawyers' effort) that they generate on behalf of clients in connection with litigation. The protection of work product is accomplished principally by denying opposing litigants the right to legal discovery of it. As Easterbrook (1981) stressed, protection of work product encourages lawyers to engage in research on their clients' cases, for much of the value of the research would be lost if it became immediately known to the other side. But whether protection of work product is socially desirable is not evident a priori; for it depends on whether or not the legal advice that the work product supports is socially desirable [155]. A further complication is that, even when the advice is socially desirable, the private value of advice, and thus the amount of work product, may be socially excessive.

Quality and truthfulness of advice. To the degree that poor or dishonest advice would be discovered and that lawyers would suffer penalties for having provided such advice, they will have reason not to do so. There are two basic types of penalty lawyers face for furnishing unsound legal advice: loss of business because of damage to reputation, and legal sanctions, in the form of damage judgments arising from malpractice actions, fines assessed by courts, or punishments imposed by professional associations. See Wilkins (1992).

5.6. *Appeals*

The appeals process – the process whereby a litigant disappointed with the decision of a first-order tribunal can seek reconsideration before a higher tribunal – is a widely observed feature of adjudication; in virtually all legal systems today, there exists a fairly general right of appeal of trial court decisions.

An important social justification for the appeals process concerns correction of error. See Shavell (1995b). Suppose that litigants possess information about the occurrence of error and that appeals courts can frequently verify it. Then litigants may be induced to bring appeals when errors are likely to have been made but not otherwise, because of the costs of appeals. This outcome may be fostered by imposition of fees for bringing appeals, so as to discourage appeal when decisions were likely to have been correct

[155] To illustrate, investigation may be necessary to determine or document facts that will improve accuracy, but the social value of greater accuracy may or may not exceed the cost of investigation. Also, with work-product protection, two parties may engage in duplicative efforts.

and thus unlikely to be reversed[156]. In other words, if there is an appropriate price for pursuing appeals, the appeals process can harness the information that litigants have about the occurrence of error and tend to remedy it.

When this process functions well, appeals not only result in error correction, they also do so cheaply, for the legal system is burdened with reconsidering only the subset of cases in which errors were more probably made. This may render society's investment in the appeals process inexpensive in comparison to the alternative it has of improving the accuracy of the trial process (by investing in the length and quality of trial court adjudication). Under that alternative approach, extra expenditure would be required in all cases rather than only in the subset of cases that are appealed. The appeals process, in other words, may be an economical way of correcting error by taking advantage of litigants' information that it has occurred[157].

5.7. Alternative dispute resolution

When parties need to resolve a dispute, they may turn not only to the state-sanctioned method of dispute resolution, namely, trial before a court, but also to arbitration and other forms of alternative dispute resolution (ADR)[158]. In examining ADR, it is helpful to distinguish between ex ante agreements to employ ADR – arrangements made before disputes arise – and ex post resort to ADR – use of ADR after disputes have arisen. See Shavell (1995a)[159].

5.7.1. Ex ante ADR agreements

Ex ante ADR agreements may be adopted because they are to the mutual benefit of the parties to a contract[160]. In particular, ADR may lower the cost of resolving disputes or reduce risk. Second, ADR may engender superior incentives for the parties through greater accuracy of results. Suppose, for instance, that substandard performance of a contract would be correctly assessed by expert arbitrators under ADR but not by courts. Then the parties to the contract might well prefer to adopt ADR because it would induce good performance, thereby raising the willingness of the promisee to pay for the contract. Third, ADR may beneficially affect the volume of adjudication. For example, it may be that the number of disputes brought under the legal process

[156] In fact, however, public fees for appeal are nominal, although private costs may be nontrivial.

[157] Daughety and Reinganum (2000b) and Spitzer and Talley (2000) consider models of the appeals process that include factors different from pure error correction.

[158] We focus on binding ADR; nonbinding ADR, such as mediation, is often used to foster settlement.

[159] On other issues raised by ADR, see Landes and R.A. Posner (1979). It should also be mentioned that there is a literature considering arbitration alone, not arbitration as an alternative to state-authorized litigation. See, for example, Ashenfelter (1989, 1992), Ashenfelter and Bloom (1984), and Farber (1980).

[160] We observe that to obtain many of the benefits noted in the text, the agreement to use ADR must be made ex ante; if the parties wait until a dispute arises, it will often be in the interest of one of the parties to refuse to accept ADR.

would be excessive, dissipating substantial resources of the parties without instigating mutually desirable changes in behavior; thus an ADR agreement that would serve to limit the number of disputes would be advantageous.

Because ex ante ADR agreements made by knowledgeable parties raise their well-being, it seems that ex ante ADR agreements should ordinarily be enforced by the legal system, as they are in fact. It is sometimes suggested that society should go further and subsidize ADR. A subsidy might be justified on second-best grounds, because the state already subsidizes ordinary litigation by not charging litigants for its full costs. It would seem, however, that the optimal solution is to remove the latter subsidy, unless it is justified on the ground of inadequate private incentives to sue.

5.7.2. Ex post ADR agreements

Parties will tend to make ex post ADR agreements in order to reduce dispute resolution costs and risk. On this account, ex post ADR would also tend to be socially desirable. A full evaluation of ex post ADR, however, must recognize other effects, notably, how the prospect that parties would adopt ADR ex post would affect their ex ante behavior. The proper analysis is similar to that bearing on the private versus the social value of settlements, in Section 5.2.3 [161].

5.8. Formulation of legal rules

Economic analysis of the operation of the legal system often takes the legal rules that are enforced as given. The formulation of legal rules itself, however, raises interesting economic issues [162]. One issue concerns the optimal level of detail of rules. On one hand, greater detail allows better-tailored control of behavior. On the other hand, greater detail involves higher compliance and litigation costs. Moreover, it cannot be assumed that parties will become informed of the precise content of more detailed rules. See Diver (1983), Ehrlich and Posner (1974), and Kaplow (1995a) [163].

Another issue is whether rules should be formulated fully ex ante, or instead should be incompletely specified initially and fully articulated only ex post, during adjudication of particular disputes. Fuller ex ante specification is more costly for the state, but may provide greater predictability for parties and hence induce better behavior, and it also may reduce adjudication costs. See Diver (1983), Ehrlich and Posner (1974), and Kaplow (1992c). Full ex ante specification of legal rules tends to be advantageous when the governed behavior is frequent and has common characteristics,

[161] Indeed, parties' adoption of ADR can be seen as a form of out-of-court settlement because use of ADR means that there will be no trial and instead the parties will be bound by the alternative they have chosen.

[162] For a survey of the literature, see Kaplow (2000).

[163] The subject of legal complexity has received particular attention in the context of the income tax. See, for example, Blumenthal and Slemrod (1992) and Kaplow (1996).

essentially because of economies of scale (the rule is formulated only once). For infrequent, heterogeneous behavior, leaving the specification of details until the stage of adjudication may save the state expense because many situations for which details may have been provided will never arise. A closely related subject is the issuance of precedents by courts; for example, major disagreements about issuing precedents concern the degree to which details of rulings beyond those necessary to decide the case before a court should be specified and when courts should take the opportunity to announce new legal rules or modify existing ones[164].

Additional issues are presented by the frequent need to modify legal rules. New rules, if fully and immediately applicable, will typically affect the returns to previous investments. The prospect of such application of new rules imposes risk on actors and also affects their investment decisions, but the latter effect tends to be efficient when the legal reform reflects certain economically relevant information or changed circumstances[165]. See generally Kaplow (1986a, 1992a).

5.9. Relevance to general incentive schemes

In closing, we suggest that the topic of the operation of the legal system should in substantial respects be viewed as a basic one in the theory of incentives. This is because incentive schemes often require that parties come before authorities who apply rules – that is, the incentive schemes – and this adjudication process is costly. (Even if the adjudication is informal, it will involve expense.) Therefore, many of the issues that we address are of relevance. Notably, questions arise concerning private versus system-appropriate motives to come before authorities who apply rules, for individuals will not take into account total adjudication costs nor the incentive effects of adjudication (such as whether the bringing of employee complaints will induce better behavior in a firm). Also, many of the more particular issues that we consider about litigation and the design of legal procedure – including settlement, discovery, and appeals – have general analogs in other incentive systems[166].

6. Law enforcement

In this section we consider the theory of public enforcement of law – the use of hired agents (inspectors, tax auditors, police) to detect violators of legal rules and

[164] See also Landes and Posner (1976), who analyze the body of precedent as a capital stock that depreciates over time.

[165] For example, suppose that the government learns that a type of emission is harmful and, accordingly, imposes some sort of regulation or corrective tax. It will tend to be efficient for the new rule to apply to preexisting sources of the emissions because the prospect of such application will induce actors to take into account the probability that the emissions will turn out to be harmful. (Compare the discussion of compensation for government takings in Section 3.4.3.)

[166] The discussion in the text concerning the cost and process of verifying variables in incentive schemes obviously bears on whether parties should include variables that analysts might treat as unverifiable but that they understand to be verifiable, to some degree of accuracy, if the parties incur sufficient costs.

to impose sanctions. Outside the scope of our discussion are many other factors that affect compliance with the law, including public programs (such as job training for low-income individuals, which affects their opportunity cost of crime) and private behavior (such as the carrying of guns and the use of locks to prevent theft)[167].

We begin by noting the justification for using public law enforcement rather than relying exclusively on private suits. Then, we analyze basic issues concerning the optimal probability, magnitude, and form of sanctions and the rule of liability. Next, we examine a variety of extensions of the central theory, including accidental harms, error, marginal deterrence, repeat offenders, self-reporting, and incapacitation. Finally, we briefly discuss criminal law in the light of the theory.

Before proceeding, we observe that economically oriented analysis of public law enforcement dates from the eighteenth-century contributions of Montesquieu (1748), Beccaria (1770), and, especially, Bentham (1789), whose investigation of deterrence was sophisticated and expansive. But, curiously, after Bentham, the subject of law enforcement lay essentially dormant in economic scholarship until the late 1960s, when Gary Becker (1968) published a highly influential article, which has led to a voluminous literature[168].

6.1. Rationale for public enforcement

A basic question is why there is a need for public enforcement of law in the light of the availability of private suits brought by victims. The answer depends very much on the locus of information about the identity of injurers. When victims of harm naturally possess knowledge of the identity of injurers, allowing private suits for damages will motivate victims to sue and thus harness the information they have for purposes of law enforcement. This may help to explain why the enforcement of contractual obligations and of accident law is primarily private[169].

When victims do not know who caused harm and penalizing wrongdoing is difficult, society tends to rely instead on public investigation and prosecution; this is broadly true of crimes and of many violations of environmental and safety regulations. Even in contexts where sanctioning violators is difficult, however, we should ask why society cannot rely on inducements to private parties – rewards of some type – to supply information or otherwise to help in sanctioning. One difficulty with such private enforcement is that if a reward is available to anyone, there might be wasteful effort devoted to finding violators (akin to excessive fishing activity). Another problem

[167] On other public policies, see, for example, Donohue and Siegelman (1998) and Wilson and Herrnstein (1985). On private behavior and crime, see, for example, Ayres and Levitt (1998), Cook et al. (1995), Lott and Mustard (1997), and Shavell (1991c).

[168] For surveys and references, see Garoupa (1997), Mookherjee (1997), and Polinsky and Shavell (2000) (which is similar to this section).

[169] It may not be the case, however, that private incentives to bring suit are optimal, as we discuss in Section 5.1.2.

is that the best technologies for finding liable parties often require coordination of many individuals, sometimes on a vast scale. Additionally, it may be advantageous for expensive information systems (fingerprint records, data banks on offenders) to be developed and maintained, even though their benefits would be hard for the private sector to capture fully; such enforcement technologies may constitute natural monopolies. An additional obstacle to private enforcement is that force (or the threat of it) may be needed to gather information, capture violators, and prevent reprisal, yet the state for various reasons may not want to permit private parties to use force. Thus, there appear to exist arguments favoring public enforcement when effort is required to identify and sanction violators[170].

6.2. Basic theory of enforcement

Suppose that an individual (or a firm) chooses whether to commit an act that causes harm with certainty (see Section 6.3.1 on uncertain harm). If he commits the act, he obtains some gain and also faces the risk of being caught, found liable, and sanctioned. The rule of liability could be either strict – under which the individual is definitely sanctioned – or fault-based – under which he is sanctioned only if his behavior fell below a fault standard[171]. The sanction that he suffers could be a monetary fine, a prison term, or a combination of the two.

Whether an individual commits a harmful act is determined by an expected utility calculation. He will commit the act if that would raise his expected utility, taking into account the gain he would derive and the probability, form, and level of sanction that he would then face[172]. We will usually first examine the case in which individuals are risk neutral with respect to sanctions, but we will also consider other possibilities.

We assume, as is conventional, that fines are socially costless to employ because they are mere transfers of money, whereas imprisonment involves positive social costs because of the expense associated with the operation of prisons and the disutility

[170] The comparison between public and private enforcement has received some, but modest, attention in the literature. See Becker and Stigler (1974), Landes and Posner (1975), and Polinsky (1980a); see also Friedman (1995) and Shavell (1993a).

[171] Fault-based liability is often employed in accident law (the negligence rule) and in many regulatory schemes (which penalize only parties that fail to meet regulatory standards). On reflection, criminal law may be seen to be fault-based; it only punishes certain harmful acts whose characteristics make them almost always undesirable.

[172] We assume that the probability and magnitude of the sanction are known. See, however, Bebchuk and Kaplow (1992) on the case where individuals misperceive the probability of sanctions, Kaplow (1990b) on the case where individuals may make expenditures to acquire information about sanctions, Garoupa (1999) on both of these cases, and Sah (1991) on the process by which individuals learn about actual levels of enforcement. For empirical evidence on offenders' knowledge of expected sanctions, see Wilson and Herrnstein (1985).

due to imprisonment (which is not naturally balanced by gains to others)[173]. We also assume that the higher is the probability of detecting and sanctioning violators, the more resources the state must devote to enforcement[174].

Social welfare generally is presumed to equal the sum of individuals' expected utilities. An individual's expected utility depends on whether he commits a harmful act, on whether he is sanctioned, on whether he is a victim of someone else's harmful act, and on his tax payment, which will reflect the costs of law enforcement, less any fine revenue collected. If individuals are risk neutral, social welfare can be expressed simply as the gains individuals obtain from committing their acts, less the harms caused, and less the costs of law enforcement. (The assumption that individuals' gains are always credited in social welfare could be relaxed without affecting most of our conclusions. The principal difference that altering the assumption would make is that more acts would be treated as socially undesirable and that optimal sanctions and enforcement effort would thus be higher.)

The enforcement authority's problem is to maximize social welfare by choosing enforcement expenditures, or, equivalently, a probability of detection, and also the level of sanctions, their form (a fine, prison term, or combination), and the rule of liability (strict or fault-based).

6.2.1. Optimal enforcement given the probability of detection

We consider here optimal enforcement given the assumption that the probability of detection is fixed. Thus, we ask about the optimal form and level of sanctions under strict and fault-based liability and about how the two liability rules compare.

Strict liability. Assume initially that fines are the form of sanction and that individuals are risk neutral. Then the optimal fine f is h/p, the harm divided by the probability of detection, for then the expected fine equals the harm. This fine is optimal because, when the expected fine equals the harm, an individual will commit a harmful act if, and only if, the gain he would derive from it exceeds the harm he would cause. Essentially this basic and fundamental formula was noted by Bentham (1789, p. 173), and it has been observed by many others since.

[173] Imposing fines may, in fact, be costly, due to the need for adjudication and fine collection. Were we to take this into account, the main effect on our conclusions would be that the optimal expected sanction would be higher because harmful acts would cause not only direct harm but also, if detected, additional administrative costs. (Note, however, that any legal costs borne by the actor are already included in his calculus, so they do not affect the optimal expected sanction.) See, for example, Becker (1968) and Polinsky and Shavell (1992).

[174] We note that when the method of enforcement involves investigating particular acts after they have been committed (rather than auditing or monitoring, such as when police walk a beat), raising the probability of apprehension may, in some ranges, involve lower costs on account of greater deterrence, which reduces the number of acts that need to be investigated to maintain a given probability of detection.

If individuals are risk averse, one might expect the optimal fine to be lower than in the risk-neutral case for two reasons. First, because risk-averse individuals are more easily deterred than risk-neutral individuals, the fine does not need to be as high as before to achieve any desired degree of deterrence. Second, lowering the fine reduces the bearing of risk by individuals who commit the harmful act. However, lowering the fine also increases the number of individuals who commit the harmful act and hence bear risk[175].

Next, assume that imprisonment is the form of sanction, so that social costs will be incurred in imposing sanctions. In this case, there is not a simple formula for the optimal imprisonment term. See Polinsky and Shavell (1984). The optimal term could be such that there is either underdeterrence or overdeterrence, compared to socially ideal behavior. On one hand, a relatively low imprisonment term, implying underdeterrence, might be socially desirable because it means that imprisonment costs are reduced for those individuals who commit harmful acts. On the other hand, a relatively high term, implying overdeterrence, might be socially desirable because it means that imprisonment costs are reduced due to fewer individuals committing harmful acts[176]. (For reasons that we will discuss below and because of factors outside the model, our conjecture is that overdeterrence is unlikely to be optimal.)

Now consider the combined use of fines and imprisonment. Here, the main point is that fines should be employed to the maximum extent feasible before resort is made to imprisonment. In other words, it is not optimal to impose a positive imprisonment term unless the fine is maximal. (The maximal fine might be interpreted as the wealth of an individual.) The rationale for this conclusion is that fines are socially costless to impose, whereas imprisonment is socially costly, so deterrence should be achieved through the cheaper form of sanction first. This point is noted by Bentham (1789, p. 183) and Becker (1968); see also Polinsky and Shavell (1984). To amplify, suppose that the fine f is less than the maximal fine f_m and that a positive prison term t is employed. Raise f toward f_m and lower t so as to keep the disutility of the combined sanctions constant. Then deterrence and the amount of harm will be unchanged, but the cost of imposing the imprisonment sanction will fall, raising social welfare. Hence, it must be optimal for the fine to be maximal before imprisonment is used[177]. (It can be shown that this argument holds regardless of individuals' attitudes toward risk in either fines or imprisonment.)

[175] An additional complication, which might favor a higher optimal fine when individuals are risk averse, is that individuals who commit a harmful act might obtain such great benefits that they would be wealthier (and thus have a lower marginal utility of wealth) than others even if they paid a fine equal to h/p. Then, raising the fine above h/p would tend to raise social welfare by transferring wealth from those who are sanctioned to others, who have a relatively higher marginal utility of wealth.

[176] See also Kaplow (1990a), who notes that extreme sanctions (zero or the maximal sanction) may well be optimal in the standard model.

[177] See Levitt (1997b) on why it may be optimal to rely more on imprisonment when offenders' wealth cannot be observed. For empirical evidence on the use of fines versus imprisonment, see Lott (1992) and Waldfogel (1995a).

Fault-based liability. As we explained in Section 2.4.1 on accident law, damages equal to harm, in excess of harm, or even somewhat less than harm, will be sufficient to induce optimal behavior under fault-based liability. The same logic is applicable here, where a sanction of h/p – implying that the expected sanction equals expected harm – plays the role of damages in our prior discussion. However, if errors occur in the legal process, deterrence may not be optimal, and excessive deterrence may result. See Section 2.1.1.

When individuals are risk averse or imprisonment is used as a sanction, fault liability has an advantage over strict liability: individuals who behave optimally, but nevertheless cause harm, will not be sanctioned. The socially costly imposition of sanctions is thus avoided. (That is, with fines, individuals who behave properly will not actually bear any risk, and with imprisonment, resources will not be wasted on such individuals.) See Shavell (1982a, 1985a, 1987b). The primary qualification to this argument is that fault-based liability is more difficult to administer because the enforcement authority must determine the fault standard and it must ascertain whether the fault standard was met. See Section 2.1.1. (Moreover, for reasons we discuss in Section 6.3.2 below, strict liability encourages better decisions by injurers regarding their level of participation in harm-creating activities.)

6.2.2. *Optimal enforcement including the probability of detection*

We now consider the optimal system of enforcement when expenditures on enforcement, and hence the probability of detection, are allowed to vary. Consideration of this issue originated with Becker (1968).

Strict liability. Assume first that the sanction is a fine and that individuals are risk neutral. Then the optimal level of the fine is maximal, f_m, and the optimal probability is low (in a sense to be described). The explanation is that if the fine were not maximal, society could save enforcement costs by simultaneously raising the fine and lowering the probability without affecting the level of deterrence: if $f < f_m$, then raise the fine to f_m and lower the probability from p to $(f/f_m)p$; the expected fine is still pf, so that deterrence is maintained, but expenditures on enforcement are reduced, implying that social welfare rises. Becker (1968) suggested this result (although much of his analysis implicitly presumes that the fine is not maximal); Carr-Hill and Stern (1979) and Polinsky and Shavell (1979) note it explicitly.

The optimal probability is low in that there is some underdeterrence: the optimal p is such that the expected fine pf_m is less than the harm h. See Polinsky and Shavell (1984). The reason for this result is that if pf_m equals h, behavior will be ideal, meaning that the individuals who are just deterred obtain gains just equal to the harm. These are the individuals who would be led to commit the harmful act if p were slightly reduced. Decreasing p, in turn, must be socially beneficial because these individuals cause no net social losses (because their gains essentially equal the harm), but reducing p saves enforcement costs.

If individuals are risk averse, the optimal fine may well be less than maximal, as shown in Polinsky and Shavell (1979). This is because the use of a very high fine would impose a substantial risk-bearing cost on individuals who commit harmful acts[178]. For further discussion of the optimal fine when individuals are risk averse, see Kaplow (1992b)[179].

Next, assume that the sanction is imprisonment and that individuals are risk neutral in imprisonment, that is, the disutility of a year of imprisonment is the same for each additional year[180]. Then the optimal imprisonment term is maximal for essentially familiar reasons: if the imprisonment term is raised and the probability of detection lowered so as to keep the expected sanction constant, neither individual behavior nor the costs of imposing imprisonment are affected (by construction, the expected prison term is the same), but enforcement expenditures fall. See Shavell (1991b). If, instead, individuals are risk averse in imprisonment (the disutility of each year of imprisonment grows with the number of years in prison), there is a stronger argument for setting the imprisonment sanction maximally than when individuals are risk neutral: when the imprisonment term is raised, the probability of detection can be lowered even more than in the risk-neutral case without reducing deterrence. Thus, not only are there greater savings in enforcement expenditures, but also the social costs of imposing imprisonment sanctions decline because the expected prison term falls. See Polinsky and Shavell (1999).

Last, suppose that individuals are risk preferring in imprisonment (the disutility of each year of imprisonment falls with the number of years in prison). This possibility seems particularly important: the first years of imprisonment may create special disutility, due to brutalization of the prisoner or due to the stigma of having been imprisoned at all, and potential offenders may have unusually high discount rates. In this case, the optimal sanction may well be less than maximal: if the sanction were raised, the probability that maintains deterrence could not be lowered proportionally, implying that the expected prison term would rise. See Polinsky and Shavell (1999).

[178] A more particular explanation involves reconsidering the argument that we used in the risk-neutral case. If the fine f is less than f_m, it is still true that f can be raised to f_m and p lowered so that prospective violators' expected utility remains constant; hence, everyone's behavior will be unchanged. However, because of risk aversion, this adjustment implies that pf falls, meaning that fine revenue falls. (The reduction in fine revenue reflects the disutility caused by imposing greater risk on risk-averse individuals.) If individuals are sufficiently risk averse, the decline in fine revenue associated with greater risk-bearing could more than offset the savings in enforcement expenditures from reducing the probability of detection, implying that taxes would have to rise to make up the shortfall; accordingly, social welfare would be lower.

[179] Another reason that optimal fines may not be maximal is that higher sanctions may induce violators to expend additional resources to avoid punishment. See Malik (1990). Further reasons are discussed below.

[180] We did not discuss individuals' attitudes toward the risk of imprisonment above because the points we made there did not depend on this consideration.

Now consider the situation when both fines and imprisonment are employed as sanctions. Recall that under the optimal enforcement policy, the fine must be maximal, for otherwise it cannot be desirable to employ imprisonment. The main point we wish to add is that, unlike when imprisonment is used alone, the optimal imprisonment term may not be maximal even if individuals are risk neutral or risk averse in imprisonment. The basic reason is that, if the imprisonment term is raised and the probability of detection is lowered so as to keep deterrence constant, there will be relatively greater reliance on imprisonment than on fines, which is more socially costly [181].

Fault-based liability. The least expensive way to accomplish compliance with the fault standard is to use the highest possible sanction and, given this sanction, the lowest probability of detection that deters individuals who would be at fault. The reason is that, if all individuals who would be at fault are deterred, the only cost incurred is associated with the setting of the probability; this cost is minimized by using the maximal sanction and a correspondingly low probability. Observe that this is true regardless of whether the sanction is a fine or imprisonment and regardless of individuals' attitudes toward the risk of fines or of imprisonment. As noted above, however, determining fault may be difficult. Errors will affect deterrence and will result in some imposition of sanctions that may be socially costly.

6.3. Extensions of the basic theory

6.3.1. Accidental harms

We initially assumed that individuals consider committing acts that cause harm with certainty. In many circumstances, however, individuals cause harms only by accident – harm occurs only with a probability. For instance, if someone drives while intoxicated, he only creates a likelihood of a collision; or if a firm stores toxic chemicals in a substandard tank, the firm only creates the probability of a harmful spill.

Essentially all that we have said above applies in a straightforward manner when harms are accidental. There is, however, an additional issue that arises when harm is uncertain: a sanction can be imposed either on the basis of the commission of a dangerous act that increases the chance of harm – storing chemicals in a substandard tank – or on the basis of the actual occurrence of harm – only if the tank ruptures and results in a spill. In principle, either approach can achieve optimal deterrence: when individuals are risk-neutral, the sanction for committing a dangerous act would equal the expected harm, and the sanction for causing harm would simply equal the magnitude of the harm itself.

Several factors are relevant to the choice between act-based and harm-based sanctions. See Shavell (1993a). First, act-based sanctions, being based only on expected

[181] Reducing the probability reduces the expected disutility attributable to fines (which are constant in nominal amount, at the maximum level); to keep deterrence constant, expected disutility attributable to imprisonment must rise. See Shavell (1991b).

harm, need not be as high to accomplish a given level of deterrence, and thus offer an underlying advantage over harm-based sanctions because of limitations in parties' assets. See Section 2.6. Such lower sanctions will also be beneficial when parties are risk averse. Second, act-based sanctions and harm-based sanctions may differ in the ease with which they can be applied. In some circumstances, act-based sanctions may be simpler to impose (it might be easier to determine whether an oil shipper properly maintains its vessels' holding tanks than to detect whether one of the vessels leaked oil into the ocean). In other circumstances, harm-based sanctions may be more readily applied (it may be easy to identify that a truck exploded but may be difficult to detect a truck illegally carrying explosives). Third, calculation of the appropriate sanction may be less difficult in one context or the other: actual harm may be apparent when it occurs, whereas the probability may be difficult to assess at the time of an act; or expected harm may be statistically determinable, but identifying actual harm (for example, tracing particular pollutants to particular victims) may be nearly impossible.

6.3.2. Level of activity

We have been assuming that the sole decision that an individual makes is whether to act in a way that causes harm when engaging in some activity. In many contexts, however, an individual also chooses whether to engage in that activity, or, more generally, at what level to do so. Thus, as we discussed in Section 2.1.3 on liability for accidents, individuals decide both how carefully to drive and how much to drive. Similarly, firms decide on a pollution technology and a level of production. And, as we observed previously, even parties who act with appropriate care may impose harm; hence, their activity levels will tend to be optimal only if they bear the cost of that residual harm. Thus, under strict liability, choices about activity levels tend to be correct, but under fault-based liability, parties generally will participate in activities to a socially excessive extent. An important application of this point concerns safety and environmental regulation. Such regulation is typically framed in terms of standards that have to be met, but which, if met, free regulated parties from liability. Under such regulation, levels of regulated activities tend to be excessive.

6.3.3. Enforcement error

Errors of the two classic types can occur in law enforcement: an individual who should be found liable might mistakenly not be found liable, and an individual who should not be found liable might mistakenly be found liable. Let the probabilities of these errors be ε_1 and ε_2, respectively, for an individual who has been detected. Thus, an individual will commit the wrongful act when his gain g net of his expected fine if he does commit it leaves him better off than paying the expected fine if he does not commit it, namely, when $g - p(1-\varepsilon_1)f > -p\varepsilon_2 f$, or, equivalently, when $g > (1-\varepsilon_1-\varepsilon_2)pf$.

The first point to note is that, as emphasized in Png (1986), both types of error reduce deterrence: the term $(1-\varepsilon_1-\varepsilon_2)pf$ is declining in both ε_1 and ε_2. The first

type of error diminishes deterrence because it lowers the expected fine if an individual violates the law. The second type of error, when an individual is mistakenly found liable, also lowers deterrence because it reduces the marginal benefit of complying with the law. Because errors dilute deterrence, they reduce social welfare. Specifically, to achieve any level of deterrence, the probability p must be higher to offset the effect of errors. Also, when sanctions are socially costly, greater sanctioning costs may be incurred to achieve a given level of deterrence[182]. See generally Kaplow and Shavell (1994a).

Now consider the optimal choice of the fine. Given any probability of detection, the dilution in deterrence caused by errors requires a higher fine to restore deterrence. If the probability and the fine are variable, then, as before, the optimal fine is maximal for the now familiar reason.

Next, consider the possible risk aversion of individuals. As we emphasized, the optimal fine under strict liability may well be less than maximal when individuals are risk averse, in part because lowering the fine from the maximum level reduces the bearing of risk. Introducing the possibility of errors may increase the desirability of lowering the fine because individuals who do not violate the law are subject to the risk of having to pay a fine[183]. Indeed, because the number of persons who do not violate the law often would far exceed the number who do, the desire to avoid imposing risk on the former group can lead to a substantial reduction in the optimal fine.

The possibility of error has analogous effects on our analysis of nonmonetary sanctions. The effect of error on the performance of fault-based liability was already noted in Section 6.2.1.

Finally, observe that, although we have treated the probabilities of error as fixed, they can be influenced by procedural choices: generally, increasing resources devoted to investigation and adjudication tends to decrease errors, and adjusting the burden of proof affects the tradeoff between the two types of errors. Because both types of error reduce deterrence and increase the imposition of socially costly sanctions for a given level of deterrence, expenditures made to reduce errors may be socially beneficial. See Kaplow and Shavell (1994a).

6.3.4. General enforcement

Enforcement sometimes is general in the sense that several different types of violations will be detected by an enforcement agent's activity. For example, a police officer waiting at the roadside may notice a driver who litters as well as one who goes through a red light or who speeds, and a tax auditor may detect a variety of infractions when he examines a tax return. To analyze such situations, suppose that a single probability

[182] First, sanctions will sometimes be imposed on those who did not commit the harmful act. Second, to maintain a given gap in disutility from sanctions due to committing the harmful act, the expected sanction for actual injurers must rise as well.

[183] See Block and Sidak (1980).

of detection applies uniformly to all harmful acts, regardless of the magnitude of the harm. (The contrasting assumption is that enforcement is specific, meaning that the probability is chosen independently for each type of harmful act[184].)

The main point that we want to make is that in contexts in which enforcement is general, the optimal sanction rises with the severity of the harm and is maximal only for relatively high harms. See Shavell (1991b); Mookherjee and Png (1992) is closely related. To explain, assume that liability is strict, the sanction is a fine, and injurers are risk neutral. Let $f(h)$ be the fine given harm h. Then, for any given general probability of detection p, the optimal fine schedule is h/p, provided that h/p is feasible; otherwise – for high h (all h such that $h/p > f_m$) – the optimal fine is maximal. This schedule is obviously optimal given p because it implies that the expected fine equals harm, thereby inducing ideal behavior, whenever that is possible.

The question remains whether it would be desirable to lower p and raise fines to the maximal level for the range of relatively low-harm acts for which h/p is less than maximal. The answer is that if p is reduced for the relatively low-harm acts (and the fine raised for them), then p, being general, is also reduced for the high-harm acts for which the fine is already maximal, which worsens the extent of underdeterrence of these acts. The decline in deterrence of high-harm acts may cause a greater social loss than the savings in enforcement costs from lowering p. To express this point differently, p must be sufficiently high to avoid significant underdeterrence of high-harm acts (for which fines are maximal). But since this p also applies to less harmful acts, the fines for them do not need to be maximal in order to deter them appropriately.

6.3.5. Marginal deterrence

Sometimes a person may consider which of several harmful acts to commit, for example, whether to release only a small amount of a pollutant into a river or a large amount, or whether only to kidnap a person or also to kill him. In such contexts, the threat of sanctions influences not only whether individuals are deterred from committing harmful acts but also, for those who are not deterred, which harmful acts they will then choose to commit. Notably, undeterred individuals will have a reason to commit less harmful rather than more harmful acts if expected sanctions rise with harm – a phenomenon that is sometimes referred to as marginal deterrence, named by Stigler (1970). The benefits of achieving marginal deterrence were noted long ago by Beccaria (1770, p. 32) and Bentham (1789, p. 171). There may, however, be costs of accomplishing marginal deterrence: for sanctions to rise with the magnitude of harm it may be necessary to apply lower sanctions to less harmful acts, which will reduce the deterrence of such acts.

[184] These assumptions correspond to different law enforcement technologies. Investigation (the police following leads after the commission of a particular crime) tends to be specific, whereas auditing and monitoring tend to be general.

Two additional observations should be made about marginal deterrence. First, marginal deterrence can be promoted by adjusting the probability of detection as well as the magnitude of sanctions. (Thus, rather than achieving marginal deterrence by lowering the sanction for the less harmful act, the state can lower the probability of detection for that act; this accomplishes the same result with regard to deterrence and saves enforcement resources[185].) Second, marginal deterrence is naturally accomplished if the expected sanction equals harm for all levels of harm; for instance, if a polluter's expected fine would rise from \$100 to \$500 if he dumps five gallons instead of one gallon of waste into a lake, where each gallon causes \$100 of harm, his marginal incentives to pollute will be correct. For formal analyses of marginal deterrence, see Friedman and Sjostrom (1993), Mookherjee and Png (1994), Shavell (1992), and Wilde (1992).

6.3.6. Repeat offenders

In practice, the law often sanctions repeat offenders more severely than first-time offenders. Note initially, however, that sanctioning repeat offenders more severely cannot be socially advantageous if deterrence always induces first-best behavior[186]. Thus, it is only when there is underdeterrence (which is often optimal even when acts are always undesirable, as with many crimes) that it might be optimal to punish repeat offenders more severely.

The main justification for a greater sanction for repeat offenders is that repeat offenders may reveal themselves to be different in some manner that bears on the optimal sanction[187]. Another reason to raise sanctions is if additional imprisonment has less deterrent effect per unit, as discussed in Section 6.2.2. We also note that greater sanctions for repeat offenders not only deter repeat offenses but also initial offenses (if these might be followed by later offenses). For analyses of repeat offenses, see Chu et al. (2000), Landsberger and Meilijson (1982), Polinsky and Rubinfeld (1991), Polinsky and Shavell (1998b), and Rubinstein (1979).

[185] When enforcement is general, the same probability will be applicable to a range of offenses, in which case adjusting sanctions may be the only way to achieve marginal deterrence.

[186] If the sanction for polluting and causing a \$1000 harm is \$1000, then any person who pollutes and pays \$1000 is a person whose gain from polluting (say, the savings from not installing pollution control equipment) must have exceeded \$1000. Social welfare therefore is higher as a result of his polluting. If such an individual polluted and was sanctioned in the past, that only means that it was socially desirable for him to have polluted previously. Raising the current sanction because of his having a record of sanctions would overdeter him now.

[187] Observe that the mere fact that the sanction for the first offense was inadequate to deter repeat offenders is not enough to justify a higher sanction, for this fact was known at the time they were sanctioned for the first offense.

6.3.7. Self-reporting

We have thus far assumed that individuals are subject to sanctions only if they are detected by an enforcement agent, but in fact parties sometimes disclose their own violations to enforcement authorities. For example, firms often are required to report their violations of environmental and safety regulations, individuals sometimes notify police of their involvement in traffic accidents, and even criminals occasionally turn themselves in. We explain here why it is generally socially desirable for the structure of enforcement to be such as to encourage self-reporting. See Kaplow and Shavell (1994b) and Malik (1993).

Self-reporting can be induced by the state lowering the sanction for individuals who disclose their own infractions. Moreover, the reward for self-reporting can be made small enough that deterrence is only negligibly reduced. To amplify, assume for simplicity that the sanction is a fine f, that the probability of detection is p, and that individuals are risk neutral. If an individual commits a violation and does not self-report, his expected fine is pf. Suppose the fine if a violator self-reports is set just below pf, say at $pf - \varepsilon$, where $\varepsilon > 0$ is arbitrarily small. Then the violator will want to self-report but the deterrent effect of the sanction will be (approximately) the same as if he did not self-report.

Given that self-reporting can be induced, essentially without compromising deterrence, why exactly is self-reporting socially advantageous? One reason is that self-reporting reduces enforcement costs: when a party self-reports, the enforcement authority does not have to identify and prove who the violator was; if a polluter or a burglar turns himself in, investigatory resources are saved[188]. Second, self-reporting reduces risk, and thus is advantageous if injurers are risk averse. Drivers bear less risk because they know that if they cause an accident, they can (and will be led to) report this to the police and suffer a lower and certain sanction, rather than face a substantially higher sanction (such as for hit and run driving) imposed only with some probability. Third, the magnitude of harm caused sometimes will be mitigated as a consequence of self-reporting; for example, when firms are induced to report leaks of toxic substances when they occur, prompt remediation is more likely to take place[189].

6.3.8. Plea bargaining

Plea bargaining refers to settlement negotiations between a public prosecutor and a criminal defendant. We examined this subject in Section 5.4.9. On plea bargaining, see, for example, Froeb (1993), Grossman and Katz (1983), Kobayashi and Lott (1996), Miceli (1996), and Reinganum (1988, 2000).

[188] When enforcement is accomplished by means of auditing or monitoring, self-reporting results in only modest savings. For self-reporting then only reduces, perhaps slightly (if most individuals comply with the law), the population of individuals to be audited or monitored.

[189] See Innes (1999).

6.3.9. Corruption of law enforcement agents

An enforcement agent and a potential violator might well find it mutually profitable to make an agreement under which the violator pays the agent to keep silent. This problem of corruption would seem to be worse the larger is the sanction faced by a violator. To combat corruption and the undermining of deterrence that it brings about, two general approaches can be employed. One is to raise the overall level of sanctions, so that bargained-for payments will also rise. This, however, is a gross strategy, and also suffers from the limit on the magnitude of sanctions that can actually be imposed. The second approach is to attempt to control corruption by use of sanctions against those who participate in it. This approach is expensive and involves issues of enforcement in its own right[190].

6.3.10. Principal–agent relationship

Although we have assumed that an injurer is a single actor, the injurer is often an agent of some principal. For example, the agent could be an employee of a firm, or the agent could be a subcontractor working for a contractor.

When harm is caused by the behavior of principals and their agents, many of the conclusions of our prior analysis carry over to the sanctioning of principals. Notably, if a risk-neutral principal faces an expected fine equal to harm done, he will in effect be in the same position vis-à-vis his agent as society is vis-à-vis a single potential violator of law. See Newman and Wright (1990). Consequently, the principal will behave socially optimally in controlling his agents and, in particular, will contract with them and monitor them in ways that will give the agents socially appropriate incentives to reduce harm[191].

A question about enforcement that arises when there are principals and agents is the allocation of financial sanctions between the two parties[192]. It is apparent, however, that the particular allocation of sanctions does not matter when, as would be the natural presumption, the parties can reallocate the sanctions through their own contract. For example, if the agent finds that he faces the risk of a large fine but is more risk averse than the principal, the principal can assume the risk; conversely, if the risk of the fine is imposed on the principal, he will retain it. Thus, the post-contract sanctions that the agent bears are not affected by the particular division of sanctions initially selected by the enforcement authority.

[190] For fairly general, mainly informal discussions of corruption, see Becker and Stigler (1974), Klitgaard (1988), Rose-Ackerman (1978, 1999), and Shleifer and Vishny (1993); and for models of various aspects of corruption, see, for example, Bowles and Garoupa (1997), Cadot (1987), Mookherjee and Png (1995), and Polinsky and Shavell (2001). In the principal–agent context, analogous problems are examined in Kofman and Lawarree (1993) and Tirole (1986).

[191] But, as Arlen (1994) indicates, firms' internal monitoring of agents might be discouraged if such monitoring makes firms' exposure to external sanctions more likely.

[192] See Kraakman (1984).

The allocation of monetary sanctions between principals and agents does matter if some allocations allow the pair to reduce their total burden. An important example is when a fine is imposed only on the agent and he is unable to pay it because his assets are less than the fine; see Kornhauser (1982) and Sykes (1981). Then, he and the principal (who often would have higher assets) would jointly escape part of the fine, diluting deterrence. Imposing the fine on the principal rather than on the agent avoids this problem[193].

A closely related point is that the imposition of imprisonment sanctions on agents may be desirable when their assets are less than the harm that they can cause, even if the principal's assets are sufficient to pay the optimal fine. See Polinsky and Shavell (1993). The fact that an agent's assets are limited means that the principal may be unable to control him adequately through use of contractually determined penalties. For example, a firm may not be able, despite the threat of salary reduction or dismissal, to induce its employees never to rig bids. In such circumstances, it may be socially valuable to use the threat of personal criminal liability and a jail sentence to improve the control of agents' misconduct.

6.3.11. Incapacitation

Our discussion of public enforcement has focused on the deterrent effect of sanctions. However, a different way for society to reduce harm is by imposing sanctions that remove parties from positions in which they are able to cause harm – that is, by incapacitating them. Imprisonment is the primary incapacitative sanction, although there are other examples: individuals can lose their drivers' licenses; businesses can lose their right to operate in certain domains, and the like. Here, we consider imprisonment, but what we say applies to incapacitative sanctions generally. On the economic theory of incapacitation, see Shavell (1987c).

To better understand the role of public enforcement when sanctions are incapacitative, suppose that the sole function of sanctions is to incapacitate; that is, sanctions do not deter. In this case, continued imprisonment will be desirable as long as the reduction in crime from incapacitation exceeds the costs of imprisonment. Observe that this condition could hold for a long period, even for offenses that are not the most serious. There is, however, evidence that the proclivity to commit crimes declines sharply with age after a certain point. We also note that, as a matter of logic, the incapacitative rationale might imply that a person should be imprisoned even if he has not committed a crime – because the danger he poses to society makes incapacitating him worthwhile. In practice, however, the fact that a person has committed a harmful act may be the best basis for predicting his future behavior, in

[193] The converse problem, when the principal has insufficient assets, may also arise. Then, it may be optimal to hold agents or other contracting parties, such as lawyers or lenders, liable as well. See, for example, Kraakman (1986) and Pitchford (1995).

which case the incapacitation rationale would suggest imprisoning an individual only if he has committed such an act[194].

Several comments may be made on the relationship between optimal enforcement when incapacitation is the goal versus when deterrence is the goal. First, when incapacitation is the goal, the optimal magnitude of the sanction is independent of the probability of apprehension, which contrasts with the case when deterrence is the goal. Second, when deterrence is the goal, the probability and magnitude of sanctions depend on the ability to deter, and if this ability is limited (as, for instance, with the insane), a low expected sanction may be optimal, whereas a high sanction still might be called for to incapacitate.

6.3.12. Empirical evidence on law enforcement

A great deal of empirical work has been devoted to controlling criminals. See, for example, Anderson (1999), Blumstein et al. (1978), Cook and Zarkin (1985), DiIulio and Piehl (1991), Ehrlich (1973, 1975), Eide (1994, 2000), Grogger (1991), Kessler and Levitt (1999), Levitt (1996, 1997a, 1998a,b), Nagin (1978), Pyle (1983), Tauchen et al. (1994), Viscusi (1986b), and Witte (1980). Some of this literature, however, does not distinguish between deterrence and incapacitation as the source of reductions in crime following from greater law enforcement. A problematic issue with which the literature grapples is the simultaneity problem. Notably, when greater law enforcement is not associated with a significant reduction in crime, the explanation could be either that deterrence and incapacitation are relatively unimportant, or else that their importance is masked because enforcement effort and sanctions are increased in response to higher crime rates.

6.4. Criminal law

The subject of criminal law may be viewed in the light of the theory of public law enforcement. See R.A. Posner (1985) and Shavell (1985a). First, that the acts in the core area of crime – robbery, murder, rape, and so forth – are punished by the sanction of imprisonment makes basic sense. Were society to rely on monetary penalties alone, deterrence of the acts in question would be grossly inadequate. Notably, the probability of sanctions for many of these acts is small, making the money sanction necessary for deterrence large, but the assets of many individuals who might commit these acts are quite limited; hence, the threat of prison is needed for deterrence. Moreover, the incapacitative aspect of imprisonment is valuable because of difficulties in deterring many of the individuals who are prone to commit criminal acts.

Second, many of the doctrines of criminal law appear to enhance social welfare. This seems true of the basic feature of criminal law that punishment is not imposed

[194] An exception may arise for certain types of mental illness, in which case we do in fact incapacitate individuals.

on all harmful acts but instead is usually confined to those that are undesirable. (For example, murder is subject to criminal sanctions, but not all accidental killing.) As we have stressed, when the socially costly sanction of imprisonment is employed, the fault system is desirable because it results in less frequent imposition of punishment than strict liability. The focus on intent in criminal law, another of its defining features, may be sensible with regard to deterrence because those who intend to do harm are more likely actually to cause harm, may be more inclined to conceal their acts, and may be harder to discourage because of the benefits they anticipate. That unsuccessful attempts to do harm are punished in criminal law is an implicit way of raising the likelihood of sanctions for undesirable acts[195]. Study of specific doctrines of criminal law seems to afford a rich opportunity for economic analysis[196].

Third, the level of sanctions commonly employed is in some respects in accord with the theory concerning optimal enforcement; notably, offenses that are relatively more serious or more difficult to detect tend to be punished more severely than others. Sanctions, however, do not always seem to be as high as the theory suggests would be optimal. To be sure, the theory surveyed above provides many reasons that optimal sanctions may be less than maximal. Yet some sanctions appear to be substantially lower than can readily be explained by the theory[197]. An important reason for this may be that the public would view as unfair the imposition of punishment that was disproportionate to the magnitude of an offense, although the precise nature of this constraint on the use of sanctions is difficult to ascertain[198].

7. Criticism of economic analysis of law

Many observers, and particularly non-economists, view economic analysis of law with skepticism. In this section, we briefly note some of the most common criticisms.

7.1. Positive analysis

It is often claimed that individuals and firms do not respond to legal rules as rational maximizers of their well-being. Sometimes this criticism of the conventional economic approach verges on an outright rejection of the use of models. Such an extreme view reflects a failure to appreciate the role of simplifying assumptions, and, accordingly,

[195] See Shavell (1990).

[196] See R.A. Posner (1985) and Shavell (1985a) on the various doctrines of criminal law, and Kaplow (1990b) and Murphy and O'Hara (1997) on whether ignorance of the law should excuse criminal liability. See also Fischel and Sykes (1996) and Khanna (1996) on corporate criminal liability.

[197] For example, fines for many traffic violations are quite low; increasing them might achieve significant savings in enforcement resources without raising other serious problems.

[198] An implication is that jurors may not always be willing to convict if sanctions are viewed as excessive. For a model of optimal enforcement in this case, see Andreoni (1991).

it can be largely dismissed. Frequently, however, the criticism is limited to particular contexts. For example, it is often asserted that decisions to commit crimes are not governed by economists' usual assumptions. Ultimately, such criticisms raise questions that can only be answered by empirical investigation.

It is also suggested that, in predicting individuals' behavior, certain standard assumptions should be modified. For example, in predicting compliance with a law, the assumption that preferences be taken as given would be inappropriate if a legal rule would change people's preferences, as some say was the case with civil rights laws. In addition, laws may frame individuals' understanding of problems, which could affect their probability assessments or willingness to pay. See, for example, Kahneman et al. (1990) on the assignment of entitlements and the Coase theorem. The emerging field of behavioral economics and work in various disciplines that address social norms is beginning to examine these sorts of issues[199].

7.2. Normative analysis

7.2.1. Distribution of income

A frequent criticism of economic analysis of law concerns its focus on efficiency, to the exclusion of the distribution of income. The claim of critics is that legal rules – such as the choice between strict liability and negligence to govern automobile–pedestrian accidents – should be selected in a manner that reflects their effects on the rich and the poor.

There is not a good reason, however, to employ legal rules to accomplish redistributive objectives given the general alternative of achieving sought-after redistribution through the income tax and transfer programs. Such direct methods of redistribution tend to be superior to redistribution through the choice of legal rules: selecting legal rules other than those that are most efficient in order to effect redistribution is itself costly, and it also will distort individuals' labor–leisure decision in the same manner as does the income tax. See Shavell (1981) and Kaplow and Shavell (1994c)[200].

Moreover, it is difficult to redistribute income systematically through the choice of legal rules. In the first place, many individuals are never involved in litigation. Also, for those who are, there is substantial income heterogeneity both among plaintiffs and

[199] See, for example, Baron (1994), Jolls et al. (1998), Kahneman et al. (1982), and Rabin (1998).

[200] There are subtle qualifications to this claim of the sort identified in the optimal income tax literature (for example, that activities that make leisure relatively more attractive than work should be disfavored), but these qualifications are largely independent of the main criticism at issue, holding that legal rules should be designed to favor whichever party has lower income. On one such qualification, see Sanchirico (2000) and Kaplow and Shavell (2000). Furthermore, this claim assumes that redistribution will have the same effect on labor effort however it is accomplished, whereas it is possible that individuals would over- or underestimate the extent of redistribution accomplished by legal rules. See Jolls (1998).

among defendants. Additionally, in contractual contexts, the choice of a legal rule often will not have any effect on distribution because contract terms, notably, the price, will adjust so that any agreement into which parties enter will continue to reflect the initial bargaining power of each party.

7.2.2. Victim compensation

Another major criticism of economic analysis of law is that it usually emphasizes the effects of legal rules on behavior, but not the compensation of victims – which, some believe, is the main purpose of private law. Economic analysis does not, though, ignore victim compensation per se; victim compensation is relevant to social welfare if victims are risk averse. However, as we have discussed, if victims can obtain insurance, as is often possible, then the legal system need not be relied on to provide compensation. Moreover, providing compensation through legal rules tends to be significantly more expensive than doing so through insurance[201].

7.2.3. Concerns for fairness

An additional source of criticism is that the welfare-economic approach slights important concerns about fairness, justice, and rights. Some of these notions refer implicitly to the appropriateness of the distribution of income and, accordingly, are encompassed by our preceding remarks. Also, to some degree, the notions are motivated by instrumental concerns. For example, the attraction of just punishment must inhere in part in its deterrent effect, and the appeal of obeying contractual promises must rest in part on the beneficial effect this has on production and exchange. To this extent, critics' concerns are already taken into account in standard welfare-economic analysis.

However, many who advance ideas of fairness and cognate notions do not regard them merely as some sort of proxy for attaining instrumental objectives. Instead, they believe that satisfying the notions is intrinsically valuable. This view too can be partially reconciled with economists' conception of social welfare: if individuals have a taste for a legal rule or institution because they regard it as fair, that should be credited in the determination of social welfare, just as any taste should. (Note that, in this case, the importance of fairness is converted from a philosophical issue to an empirical question about individuals' tastes.)

But many uphold the view that notions of fairness are important as ethical principles in themselves, without regard to any possible relationship the principles may have to individuals' welfare. This opinion is, of course, the subject of longstanding debate

[201] The reader will also recall from Section 2.2 that there may be additional reasons that providing compensation through the legal system is undesirable, including adverse effects of victims' incentives and imposing risk on injurers.

among moral philosophers[202]. Some readers (along with us) may be skeptical of normative views that are not grounded in individuals' well-being because embracing such views entails a willingness to sacrifice individuals' well-being. Indeed, consistently pursuing any non-welfarist principle will sometimes result in everyone being made worse off[203]. Nevertheless, it is clear that such views will be reflected in criticism of economic analysis of law for the foreseeable future[204].

7.3. Purported efficiency of judge-made law

Also criticized is the contention of some economically-oriented academics – notably, R.A. Posner (1975) and Landes and Posner (1987a) – that judge-made law tends to be efficient (in contrast to legislation, which is said to reflect the influence of special interest groups)[205]. Instead, critics believe that the judge-made law is guided by notions of fairness, justice, and rights, and thus will not necessarily be efficient. Several observations about these competing views may be made. First, one would certainly expect legal rules to promote efficiency, at least in a very approximate sense, for that is consistent with many notions of fairness and with common sense. But second, one would not expect that legal rules would be efficient in a detailed sense for a variety of reasons[206]. Third, we note that judge-made law is peculiar to "common-law" countries, those of the former British Commonwealth, yet common-law legal rules are not markedly different from those in the civil-law countries of Continental Europe, which rely more on statutes and less on judicial development than common-law countries. Moreover, to the extent that legal rules in common-law and civil-law systems differ, it is hardly clear that the typical civil-law rules are less efficient[207]. Finally, it should be emphasized that the economic efficiency thesis is a particular descriptive claim about the law, and its validity does not bear on the power of

[202] We note, however, that much of the philosophical debate is about what principles should guide personal behavior in everyday life, which may not be applicable to the determination of what principles should guide social policy. A related distinction is emphasized in Hare (1981).

[203] See Kaplow and Shavell (2001).

[204] Kaplow and Shavell (2002a) provides an extensive investigation of the issues discussed in this section. See generally Sen and Williams (1982) for representative views of leading economists and philosophers on normative analysis.

[205] The argument is advanced by examining particular common-law rules and presenting arguments that they are efficient. In addition, the argument has been examined in the context of models of common-law evolution. See Cooter and Kornhauser (1980), Priest (1977), and Rubin (1977).

[206] Legal rules are arguably influenced by notions of fairness, which only loosely reflect instrumental objectives, and are also determined by a multiplicity of institutional and historically contingent factors. Moreover, even if lawmakers were attempting to promote efficiency over the course of history, they would have had a limited understanding of the relevant theory and little empirical evidence to guide them.

[207] See, for example, Shavell (1987a), who makes some comparisons of typical common-law and civil-law rules on accident law.

economics to predict behavior in response to legal rules or on the merits of normative economic analysis of law.

8. Conclusion

Having surveyed the basic areas of economic analysis of law, let us comment on possible directions for future research. Although accident liability has been fairly well explored, relatively little formal work has been done on the subject of property law. With regard to contract law, most analysis has concerned remedies for breach, but little attention has been paid to contract formation. In the area of litigation, research effort so far has focused on settlement versus trial, whereas other aspects of litigation, including its adversarial character and its optimal design, merit study. With regard to law enforcement, an issue worthy of further consideration is the incentives of enforcers (including the problem of corruption); also, many of the doctrines of criminal law deserve investigation.

Moreover, there is a very general need for empirical work on the legal system to be undertaken. One area of study is suggested by the fact that, as we emphasized, the private and the social incentives to use the legal system can be expected to diverge. Consequently, society needs estimates of the benefits and costs of legal activity in broad domains (such as auto accidents, product liability) in order to devise appropriate policy. Another potential research area is investigation of the provisions in actual contracts in various settings, in order to test and to inform the extensive theoretical work in the field. An additional subject that seems ripe for empirical study is the litigation process, especially the determinants of suit and settlement decisions and the effects of procedures for the conduct of discovery and trial. Overall, in the areas of law considered in this survey, relatively few of the particular topics that were covered have been the subject of serious empirical work, and the opportunity for progress appears to be substantial.

Acknowledgments

We thank Alan Auerbach, Steven Levitt, A. Mitchell Polinsky, Kathryn Spier, and Tanguy van Ypersele for comments, Judson Berkey, Jerry Fang, Bo Li, Eric Selmon, and Chad Shirley for research assistance, and the John M. Olin Center for Law, Economics, and Business at Harvard Law School for financial support.

References

Administrative Office of the United States Courts (1995), "Judicial business of the United States Courts", Report of the Director (Administrative Office of the United States Courts, Washington, D.C.).

Aghion, P., and P. Bolton (1987), "Contracts as a barrier to" entry", American Economic Review 77:388–401.
Aghion, P., and B.E. Hermalin (1990), "Legal restrictions on private contracts can enhance efficiency", Journal of Law, Economics, & Organization 6:381–409.
Aghion, P., M. Dewatripont and P. Rey (1994), "Renegotiation design with unverifiable information", Econometrica 62:257–282.
Alchian, A.A., and H. Demsetz (1972), "Production, information costs, and economic organization", American Economic Review 62:777–795.
Alston, L.J., G.D. Libecap and R. Schneider (1996), "The determinants and impact of property rights: land titles on the Brazilian frontier", Journal of Law, Economics, & Organization 12:25–61.
Anderson, D.A. (1999), "The aggregate burden of crime", Journal of Law and Economics 42:611–642.
Andreoni, J. (1991), "Reasonable doubt and the optimal magnitude of fines: should the penalty fit the crime?" Rand Journal of Economics 22:385–395.
Arlen, J. (1994), "The potentially perverse effects of corporate criminal liability", Journal of Legal Studies 23:833–867.
Ashenfelter, O. (1989), "Evidence on U.S. experiences with dispute resolution systems", in: W.-C. Huang, ed., Organized Labor at the Crossroads (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI) pp. 139–162.
Ashenfelter, O. (1992), "An experimental comparison of dispute rates in alternative arbitration systems", Econometrica 60:1407–1433.
Ashenfelter, O., and D.E. Bloom (1984), "Models of arbitrator behavior: theory and evidence", American Economic Review 74:111–124.
Atkinson, A.B. (1976), "The income tax treatment of charitable contributions", in: R.E. Grieson, ed., Public and Urban Economics: Essays in Honor of William S. Vickrey (D.C. Heath, Lexington, MA) pp. 13–29.
Atwood, D.A. (1990), "Land registration in Africa: the impact on agricultural production", World Development 18:659–671.
Ayres, I., and R. Gertner (1989), "Filling gaps in incomplete contracts: an economic theory of default rules", Yale Law Journal 99:87–130.
Ayres, I., and S.D. Levitt (1998), "Measuring positive externalities from unobservable victim precaution: an empirical analysis of Lojack", Quarterly Journal of Economics 113:43–77.
Baron, J. (1994), Thinking and Deciding (Cambridge University Press, New York).
Barton, J.H. (1972), "The economic basis of damages for breach of contract", Journal of Legal Studies 1:277–304.
Bebchuk, L.A. (1984), "Litigation and settlement under imperfect information", Rand Journal of Economics 15:404–415.
Bebchuk, L.A. (1988), "Suing solely to extract a settlement offer", Journal of Legal Studies 17:437–450.
Bebchuk, L.A. (1996), "A new theory concerning the credibility and success of threats to sue", Journal of Legal Studies 25:1–25.
Bebchuk, L.A., and O. Ben-Shahar (2001), "Precontractual reliance", Journal of Legal Studies 30: 423–457.
Bebchuk, L.A., and H.F. Chang (1996), "An analysis of fee shifting based on the margin of victory: on frivolous suits, meritorious suits, and the role of rule 11", Journal of Legal Studies 25:371–403.
Bebchuk, L.A., and H.F. Chang (1999), "The effect of offer-of-settlement rules on the terms of settlement", Journal of Legal Studies 28:489–513.
Bebchuk, L.A., and L. Kaplow (1992), "Optimal sanctions when individuals are imperfectly informed about the probability of apprehension", Journal of Legal Studies 21:365–370.
Bebchuk, L.A., and I.P.L. Png (1999), "Damage measures for inadvertent breach of contract", International Review of Law and Economics 19:319–331.
Bebchuk, L.A., and S. Shavell (1991), "Information and the scope of liability for breach of contract: the rule of Hadley v. Baxendale", Journal of Law, Economics, & Organization 7:284–312.

Beccaria, C. (1872), An Essay on Crimes and Punishments (W.C. Little, Albany); originally published 1770.

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

Becker, G.S., and G.J. Stigler (1974), "Law enforcement, malfeasance, and compensation of enforcers", Journal of Legal Studies 3:1–18.

Bentham, J. (1789), "An introduction to the principles of morals and legislation", in: The Utilitarians, 1973 edition (Anchor Books, Garden City, NY) pp. 5–398.

Bentham, J. (1827), Rationale of Judicial Evidence, Vol. 5 (Hunt and Clarke, London).

Bentham, J. (1830), The Theory of Legislation, 1931, edited by C.K. Ogden, translated by R. Hildreth (Kegan Paul and Co., London).

Bernstein, L. (1992), "Opting out of the legal system: extralegal contractual relations in the diamond industry", Journal of Legal Studies 21:115–157.

Bernstein, L. (1998), "Private commercial law", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 3 (Macmillan, London) pp. 108–114.

Besen, S.M., and L.J. Raskind (1991), "An introduction to the law and economics of intellectual property", Journal of Economic Perspectives 5(1):3–27.

Besley, T. (1995), "Property rights and investment incentives: theory and evidence from Ghana", Journal of Political Economy 103:903–937.

Besley, T. (1998), "Investment incentives and property rights", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 2 (Macmillan, London) pp. 359–365.

Biblowit, C. (1991), "International law and the allocation of property rights in common resources", New York International Law Review 4:77–85.

Birmingham, R.L. (1970), "Breach of contract, damage measures, and economic efficiency", Rutgers Law Review 24:273–292.

Bishop, W. (1985), "The choice of remedy for breach of contract", Journal of Legal Studies 14:299–320.

Blackstone, W. (1765–1769), Commentaries on the Laws of England; 1992 reprint of 1st edition (William S. Hein and Co., Buffalo).

Block, M.K., and J.G. Sidak (1980), "The cost of antitrust deterrence: Why not hang a price fixer now and then?", Georgetown Law Journal 68:1131–1139.

Blume, L., D.L. Rubinfeld and P. Shapiro (1984), "The taking of land: when should compensation be paid?" Quarterly Journal of Economics 99:71–92.

Blumenthal, M., and J. Slemrod (1992), "The compliance cost of the U.S. individual income tax system: a second look after tax reform", National Tax Journal 45:185–202.

Blumstein, A., J. Cohen and D. Nagin, eds (1978), Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates (National Academy of Science, Washington, D.C.)

Bouckaert, B., and G. De Geest (1995), "Private takings, private taxes, private compulsory services: the economic doctrine of quasi contracts", International Review of Law and Economics 15:463–487.

Bouckaert, B., and G. De Geest, eds (2000), Encyclopedia of Law and Economics, Vols. I–V (Edward Elgar, Cheltenham, UK).

Bovenberg, A.L., and L.H. Goulder (2002), "Environmental taxation and regulation", in: A.J. Auerbach and M. Feldstein, eds., Handbook of Public Economics, Vol. 3 (North-Holland, Amsterdam) ch. 23, this volume.

Bowles, R., and N. Garoupa (1997), "Casual police corruption and the economics of crime", International Review of Law and Economics 17:75–87.

Braeutigam, R., B. Owen and J. Panzar (1984), "An economic analysis of alternative fee shifting systems", Law and Contemporary Problems 47:173–185.

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

Bundy, S.McG., and E.R. Elhauge (1991), "Do lawyers improve the adversary system? A general theory of litigation advice and its regulation", California Law Review 79:313–420.

Bundy, S.McG., and E.R. Elhauge (1993), "Knowledge about legal sanctions", Michigan Law Review 92:261–335.
Cadot, O. (1987), "Corruption as a gamble", Journal of Public Economics 33:223–244.
Calabresi, G. (1970), The Costs of Accidents (Yale University Press, New Haven).
Calabresi, G. (1975), "Concerning cause and the law of torts", University of Chicago Law Review 43:69–108.
Calabresi, G., and A.D. Melamed (1972), "Property rules, liability rules, and inalienability: one view of the cathedral", Harvard Law Review 85:1089–1128.
Carr-Hill, R.A., and N.H. Stern (1979), Crime, the Police and Criminal Statistics: An Analysis of Official Statistics for England and Wales Using Econometric Methods (Academic Press, London).
Chang, H.F. (1995), "Patent scope, antitrust policy, and cumulative innovation", Rand Journal of Economics 26:34–57.
Charny, D. (1990), "Nonlegal sanctions in commercial relationships", Harvard Law Review 104:373–467.
Che, Y.-K. (1996), "Equilibrium formation of class action suits", Journal of Public Economics 62: 339–361.
Che, Y.-K., and T.-Y. Chung (1999), "Contract damages and cooperative investments", Rand Journal of Economics 30:84–105.
Che, Y.-K., and D.B. Hausch (1999), "Cooperative investments and the value of contracting", American Economic Review 89:125–147.
Che, Y.-K., and J.G. Yi (1993), "The role of precedents in repeated litigation", Journal of Law, Economics, & Organization 9:399–424.
Chu, C.Y.C., S.-C. Hu and T.-Y. Huang (2000), "Punishing repeat offenders more severely", International Review of Law and Economics 20:127–140.
Chung, T.-Y. (1991), "Incomplete contracts, specific investments, and risk sharing", Review of Economic Studies 58:1031–1042.
Chung, T.-Y. (1992), "On the social optimality of liquidated damage clauses: an economic analysis", Journal of Law, Economics, & Organization 8:280–305.
Coase, R.H. (1937), "The nature of the firm", Economica 4:386–405.
Coase, R.H. (1960), "The problem of social cost", Journal of Law and Economics 3:1–44.
Coffee, J.C. (1986), "Understanding the plaintiffs' attorney: the implications of economic theory for private enforcement of law through class and derivative actions", Columbia Law Review 86:669–727.
Cohen, M.A. (1992), "The motives of judges: empirical evidence from antitrust sentencing", International Review of Law and Economics 12:13–30.
Cook, P.J., and G. Zarkin (1985), "Crime and the business cycle", Journal of Legal Studies 14:115–128.
Cook, P.J., S. Molliconi and T.B. Cole (1995), "Regulating gun markets", Journal of Criminal Law and Criminology 86:59–92.
Cooter, R.D. (1985), "Unity in tort, contract, and property: the model of precaution", California Law Review 73:1–51.
Cooter, R.D. (1989), "Punitive damages for deterrence: when and how much?" Alabama Law Review 40:1143–1196.
Cooter, R.D., and L.A. Kornhauser (1980), "Can litigation improve the law without the help of judges?" Journal of Legal Studies 9:139–163.
Cooter, R.D., and D.L. Rubinfeld (1989), "Economic analysis of legal disputes and their resolution", Journal of Economic Literature 27:1067–1097.
Cooter, R.D., and D.L. Rubinfeld (1994), "An economic model of legal discovery", Journal of Legal Studies 23:435–463.
Cooter, R.D., and T. Ulen (1997), Law and Economics, 2nd edition (Addison-Wesley, Reading, MA).
Craswell, R. (1988), "Contract remedies, renegotiation, and the theory of efficient breach", Southern California Law Review 61:629–670.
Craswell, R. (1996), "Offer, acceptance, and efficient reliance", Stanford Law Review 48:481–553.

Craswell, R., and J.E. Calfee (1986), "Deterrence and uncertain legal standards", Journal of Law, Economics, & Organization 2:279–303.
Croson, R., and J.S. Johnston (2000), "Experimental results on bargaining under alternative property rights regimes", Journal of Law, Economics, & Organization 16:50–73.
Dam, K.W. (1975), "Class actions: efficiency, compensation, deterrence, and conflict of interest", Journal of Legal Studies 4:47–73.
Dana, J.D., and K.E. Spier (1993), "Expertise and contingent fees: the role of asymmetric information in attorney compensation", Journal of Law, Economics, & Organization 9:349–367.
Danzon, P.M. (1983), "Contingent fees for personal injury litigation", Bell Journal of Economics 14:213–224.
Danzon, P.M. (1985), Medical Malpractice: Theory, Evidence, and Public Policy (Harvard University Press, Cambridge, MA).
Daughety, A.F. (2000), "Settlement", in: B. Bouckaert and G. De Geest, eds., Encyclopedia of Law and Economics, Vol. V (Edward Elgar, Cheltenham, UK) pp. 95–158.
Daughety, A.F., and J.F. Reinganum (1993), "Endogenous sequencing in models of settlement and litigation", Journal of Law, Economics, & Organization 9:314–348.
Daughety, A.F., and J.F. Reinganum (1994), "Settlement negotiations with two-sided asymmetric information: model duality, information distribution, and efficiency", International Review of Law and Economics 14:283–298.
Daughety, A.F., and J.F. Reinganum (1995), "Keeping society in the dark: on the admissibility of pretrial negotiations as evidence in court", Rand Journal of Economics 26:203–221.
Daughety, A.F., and J.F. Reinganum (2000a), "On the economics of trials: adversarial process, evidence, and equilibrium bias", Journal of Law, Economics, & Organization 16:365–394.
Daughety, A.F., and J.F. Reinganum (2000b), "Appealing judgments", Rand Journal of Economics 31:502–525.
Davis, M.L. (1994), "The value of truth and the optimal standard of proof in legal disputes", Journal of Law, Economics, & Organization 10:343–359.
Demsetz, H. (1967), "Toward a theory of property rights", American Economic Association Papers and Proceedings 57:347–359.
Derfner, M.F., and A.D. Wolf (1995), Court Awarded Attorney Fees (Matthew Bender, New York).
DeVany, A.S., R.D. Eckert, C.J. Meyers, D.J. O'Hara and R.C. Scott (1969), "A property system for market allocation of the electromagnetic spectrum: a legal-economic-engineering study", Stanford Law Review 21:1499–1561.
Devlin, R.A. (1990), "Some welfare implications of no-fault automobile insurance", International Review of Law and Economics 10:193–205.
Dewatripont, M., and J. Tirole (1999), "Advocates", Journal of Political Economy 107:1–39.
Dewees, D., D. Duff and M.J. Trebilcock (1996), Exploring the Domain of Accident Law: Taking the Facts Seriously (Oxford University Press, New York).
Diamond, P.A. (1974), "Single activity accidents", Journal of Legal Studies 3:107–164.
Diamond, P.A. (1997), "Efficiency effects of punitive damages", Working Paper 97-17 (MIT Department of Economics).
Diamond, P.A., and E. Maskin (1979), "An equilibrium analysis of search and breach of contract, I: steady states", Bell Journal of Economics 10:282–316.
DiIulio, J.J., and A.M. Piehl (1991), "Does prison pay?: The stormy national debate over the cost-effectiveness of imprisonment", Brookings Review 9(4):28–35.
Diver, C.S. (1983), "The optimal precision of administrative rules", Yale Law Journal 93:65–109.
Donohue, J.J., and P. Siegelman (1998), "Allocating resources among prisons and social programs in the battle against crime", Journal of Legal Studies 27:1–43.
Easterbrook, F.H. (1981), "Insider trading, secret agents, evidentiary privileges, and the production of information", Supreme Court Review 1981:309–365.

Easterbrook, F.H., W.M. Landes and R.A. Posner (1980), "Contribution among antitrust defendants: a legal and economic analysis", Journal of Law and Economics 23:331–370.
Eckert, R.D. (1979), The Enclosure of Ocean Resources: Economics and the Law of the Sea (Hoover Institution, Stanford).
Edlin, A.S. (1996), "Cadillac contracts and up-front payments: efficient investment under expectation damages", Journal of Law, Economics, & Organization 12:98–118.
Edlin, A.S., and S. Reichelstein (1996), "Holdups, standard breach remedies, and optimal investment", American Economic Review 86:478–501.
Ehrlich, I. (1973), "Participation in illegitimate activities: a theoretical and empirical investigation", Journal of Political Economy 81:521–565.
Ehrlich, I. (1975), "The deterrent effect of capital punishment: a question of life and death", American Economic Review 65:397–417.
Ehrlich, I., and R.A. Posner (1974), "An economic analysis of legal rulemaking", Journal of Legal Studies 3:257–286.
Eide, E. (1994), Economics of Crime: Deterrence and the Rational Offender (Elsevier, New York).
Eide, E. (2000), "Economics of criminal behavior", in: B. Bouckaert and G. De Geest, eds., Encyclopedia of Law and Economics, Vol. V (Edward Elgar, Cheltenham, UK) pp. 345–389.
Eisenberg, T. (1990), "Testing the selection effect: a new theoretical framework with empirical tests", Journal of Legal Studies 19:337–358.
Eisenberg, T., and H.S. Farber (1997), "The litigious plaintiff hypothesis: case selection and resolution", Rand Journal of Economics 28:S92–S112.
Eisenberg, T., J. Goerdt, B.J. Ostrom, D. Rottman and M.T. Wells (1997), "The predictability of punitive damages", Journal of Legal Studies 26:623–661.
Elder, H.W. (1987), "Property rights structures and criminal courts: an analysis of state criminal courts", International Review of Law and Economics 7:21–32.
Ellickson, R.C. (1989), "A hypothesis of wealth-maximizing norms: evidence from the whaling industry", Journal of Law, Economics, & Organization 5:83–97.
Ellickson, R.C. (1991), Order Without Law: How Neighbors Settle Disputes (Harvard University Press, Cambridge, MA).
Ellickson, R.C. (1993), "Property in land", Yale Law Journal 102:1315–1400.
Emons, W. (1990), "Efficient liability rules for an economy with non-identical individuals", Journal of Public Economics 42:89–104.
Emons, W. (2000), "Expertise, contingent fees, and insufficient attorney effort", International Review of Law and Economics 20:21–33.
Emons, W., and J. Sobel (1991), "On the effectiveness of liability rules when agents are not identical", Review of Economic Studies 58:375–390.
Farber, H.S. (1980), "An analysis of final-offer arbitration", Journal of Conflict Resolution 24:683–705.
Farber, H.S., and M.J. White (1991), "Medical malpractice: an empirical examination of the litigation process", Rand Journal of Economics 22:199–217.
Farmer, A., and P. Pecorino (1996), "Issues of informational asymmetry in legal bargaining", in: D.A. Anderson, ed., Dispute Resolution: Bridging the Settlement Gap (JAI Press, Greenwich, CT) pp. 79–105.
Farnsworth, E.A. (1982), Contracts (Little, Brown, Boston).
Feder, G., and D. Feeny (1991), "Land tenure and property rights: theory and implications for development policy", World Bank Economic Review 5:135–153.
Feess, E. (1999), "Lender liability for environmental harm: an argument against negligence based rules", European Journal of Law and Economics 8:231–250.
Fischel, D.R. (1998), "Lawyers and confidentiality", University of Chicago Law Review 65:1–33.
Fischel, D.R., and A.O. Sykes (1996), "Corporate crime", Journal of Legal Studies 25:319–349.
Fischel, W.A. (1995), Regulatory Takings: Law, Economics, and Politics (Harvard University Press, Cambridge, MA).

Fishman, M.J., and K.M. Hagerty (1990), "The optimal amount of discretion to allow in disclosure", Quarterly Journal of Economics 105:427–444.
Friedman, A.E. (1969), "An analysis of settlement", Stanford Law Review 22:67–100.
Friedman, D.D. (1988), "Does altruism produce efficient outcomes? Marshall versus Kaldor", Journal of Legal Studies 17:1–13.
Friedman, D.D. (1995), "Making sense of English law enforcement in the eighteenth century", The University of Chicago Law School Roundtable 2:475–505.
Friedman, D.D., and W. Sjostrom (1993), "Hanged for a sheep – the economics of marginal deterrence", Journal of Legal Studies 22:345–366.
Friedman, D.D., W.M. Landes and R.A. Posner (1991), "Some economics of trade secret law", Journal of Economic Perspectives 5(1):61–72.
Froeb, L.M. (1993), "The adverse selection of cases for trial", International Review in Law and Economics 13:317–324.
Froeb, L.M., and B.H. Kobayashi (1996), "Naive, biased, yet Bayesian: can juries interpret selectively produced evidence?" Journal of Law, Economics, & Organization 12:257–276.
Fudenberg, D., and J. Tirole (1990), "Moral hazard and renegotiation in agency contracts", Econometrica 58:1279–1319.
Gambetta, D. (1993), The Sicilian Mafia: The Business of Private Protection (Harvard University Press, Cambridge, MA).
Garoupa, N. (1997), "The theory of optimal law enforcement", Journal of Economic Surveys 11:267–295.
Garoupa, N. (1999), "Optimal law enforcement with dissemination of information", European Journal of Law and Economics 7:183–196.
Gilbert, R., and C. Shapiro (1990), "Optimal patent length and breadth", Rand Journal of Economics 21:106–112.
Ginsburg, D.H., and P. Shechtman (1993), "Blackmail: an economic analysis of the law", University of Pennsylvania Law Review 141:1849–1876.
Goetz, C.J., and R.E. Scott (1980), "Enforcing promises: an examination of the basis of contract", Yale Law Journal 89:1261–1322.
Goldberg, V.P. (1974), "The economics of product safety and imperfect information", Bell Journal of Economics 5:683–688.
Gordon, H.S. (1954), "The economic theory of a common property resource: the fishery", Journal of Political Economy 62:124–142.
Gordon, W.J., and R.G. Bone (2000), "Copyright", in: B. Bouckaert and G. De Geest, eds., Encyclopedia of Law and Economics, Vol. II (Edward Elgar, Cheltenham, UK) pp. 189–215.
Gould, J.P. (1973), "The economics of legal conflicts", Journal of Legal Studies 2:279–300.
Grady, M.F. (1983), "A new positive economic theory of negligence", Yale Law Journal 92:799–829.
Gravelle, H.S.E. (1993), "The efficiency implications of cost-shifting rules", International Review of Law and Economics 13:3–18.
Green, J. (1976), "On the optimal structure of liability laws", Bell Journal of Economics 7:553–574.
Green, J., and S. Scotchmer (1995), "On the division of profit in sequential innovation", Rand Journal of Economics 26:20–33.
Greif, A. (1998), "Informal contract enforcement: lessons from medieval trade", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 2 (Macmillan, London) pp. 287–295.
Grogger, J. (1991), "Certainty vs. severity of punishment", Economic Inquiry 29:297–309.
Grossman, G., and M.L. Katz (1983), "Plea bargaining and social welfare", American Economic Review 73:749–757.
Grossman, S.J. (1981), "The informational role of warranties and private disclosure about product quality", Journal of Law and Economics 24:461–483.
Grossman, S.J., and O.D. Hart (1986), "The costs and benefits of ownership: a theory of vertical and lateral integration", Journal of Political Economy 94:691–719.

Grout, P.A. (1984), "Investment and wages in the absence of binding contracts: a Nash bargaining approach", Econometrica 52:449–460.
Hadfield, G.K. (1994), "Judicial competence and the interpretation of incomplete contracts", Journal of Legal Studies 23:159–184.
Hansmann, H. (1996), The Ownership of Enterprise (Harvard University Press, Cambridge, MA).
Hansmann, H., and R.H. Kraakman (1991), "Toward unlimited shareholder liability for corporate torts", Yale Law Journal 100:1879–1934.
Hardin, G. (1968), "The tragedy of the commons", Science 162:1243–1248.
Hare, R.M. (1981), Moral Thinking: Its Levels, Method and Point (Oxford University Press, New York).
Hart, O.D. (1987), "Incomplete contracts", in: J. Eatwell, M. Milgate, and P. Newman, eds., The New Palgrave Dictionary of Economics, Vol. 13 (Macmillan Press, New York) pp. 752–759.
Hart, O.D. (1989), "An economist's perspective on the theory of the firm", Columbia Law Review 89:1757–1774.
Hart, O.D. (1995), Firms, Contracts, and Financial Structure (Oxford University Press, New York).
Hart, O.D., and B.R. Holmström (1987), "The theory of contracts", in: T.F. Bewley, ed., Advances in Economic Theory: Fifth World Congress (Cambridge University Press, New York) pp. 71–155.
Hart, O.D., and J. Moore (1988), "Incomplete contracts and renegotiation", Econometrica 56:755–785.
Hart, O.D., and J. Moore (1990), "Property rights and the nature of the firm", Journal of Political Economy 98:1119–1158.
Hart, O.D., A. Shleifer and R.W. Vishny (1997), "The proper scope of government: theory and an application to prisons", Quarterly Journal of Economics 112:1127–1161.
Hause, J.C. (1989), "Indemnity, settlement, and litigation, or I'll be suing you", Journal of Legal Studies 18:157–179.
Hay, B.L. (1994), "Civil discovery: its effects and optimal scope", Journal of Legal Studies 23:481–515.
Hay, B.L. (1995), "Effort, information, settlement, trial", Journal of Legal Studies 24:29–62.
Hay, B.L. (1996), "Contingent fees and agency costs", Journal of Legal Studies 25:503–533.
Hay, B.L. (1997), "Optimal contingent fees in a world of settlement", Journal of Legal Studies 26:259–278.
Hay, B.L., and K.E. Spier (1997), "Burdens of proof in civil litigation: an economic perspective", Journal of Legal Studies 26:413–431.
Hay, B.L., and K.E. Spier (1998), "Settlement of litigation", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 3 (Macmillan, London) pp. 442–451.
Heller, M.A., and R.S. Eisenberg (1998), "Can patents deter innovation? The anticommons in biomedical research", Science 280:698–701.
Hermalin, B.E., and M.L. Katz (1993), "Judicial modification of contracts between sophisticated parties: a more complete view of incomplete contracts and their breach", Journal of Law, Economics, & Organization 9:230–255.
Higgins, R.S. (1978), "Producers' liability and product-related accidents", Journal of Legal Studies 7:299–321.
Higgins, R.S., and P.H. Rubin (1980), "Judicial discretion", Journal of Legal Studies 9:129–138.
Hirshleifer, J. (1971), "The private and social value of information and the reward to inventive activity", American Economic Review 61:561–574.
Hobbes, T. (1651), Leviathan, Parts One and Two, 1958 ed. by H.W. Schneider (Liberal Arts Press, New York).
Hoffman, E., and M. Spitzer (1982), "The Coase theorem: some experimental tests", Journal of Law and Economics 25:73–98.
Holmström, B.R., and J. Tirole (1989), "The theory of the firm", in: R. Schmalensee and R.D. Willig, eds., Handbook of Industrial Organization, Vol. 1 (Elsevier, New York) pp. 61–133.
Hughes, J.W., and E.A. Snyder (1995), "Litigation and settlement under the English and American rules: theory and evidence", Journal of Law and Economics 38:225–250.

Hylton, K.N. (1990), "The influence of litigation costs on deterrence under strict liability and under negligence", International Review of Law and Economics 10:161–171.
Hylton, K.N. (1993), "Asymmetric information and the selection of disputes for litigation", Journal of Legal Studies 22:187–210.
Innes, R. (1999), "Remediation and self-reporting in optimal law enforcement", Journal of Public Economics 72:379–393.
Jensen, M., and W.H. Meckling (1976), "Theory of the firm: managerial behavior, agency costs, and capital structure", Journal of Financial Economics 3:305–360.
Jolls, C. (1997), "Contracts as bilateral commitments: a new perspective on contract modification", Journal of Legal Studies 26:203–237.
Jolls, C. (1998), "Behavioral economic analysis of redistributive legal rules", Vanderbilt Law Review 51:1653–1677.
Jolls, C., C.R. Sunstein and R.H. Thaler (1998), "A behavioral approach to law and economics", Stanford Law Review 50:1471–1550.
Joskow, P.L. (1977), "Commercial impossibility, the uranium market and the Westinghouse case", Journal of Legal Studies 6:119–176.
Kahan, M. (1989), "Causation and incentives to take care under the negligence rule", Journal of Legal Studies 18:427–447.
Kahneman, D., P. Slovic and A. Tversky, eds (1982), Judgment Under Uncertainty: Heuristics and Biases (Cambridge University Press, New York).
Kahneman, D., J.L. Knetsch and R.H. Thaler (1990), "Experimental tests of the endowment effect and the Coase theorem", Journal of Political Economy 98:1325–1348.
Kakalik, J.S., P.A. Ebener, W.L.F. Felstiner and M.G. Shanley (1983), "Costs of asbestos litigation", Report R-3042-ICJ (Rand Corporation, Santa Monica, CA).
Kaplow, L. (1984), "The patent–antitrust intersection: a reappraisal", Harvard Law Review 97:1813–1892.
Kaplow, L. (1986a), "An economic analysis of legal transitions", Harvard Law Review 99:509–617.
Kaplow, L. (1986b), "Private versus social costs in bringing suit", Journal of Legal Studies 15:371–385.
Kaplow, L. (1990a), "A note on the optimal use of nonmonetary sanctions", Journal of Public Economics 42:245–247.
Kaplow, L. (1990b), "Optimal deterrence, uninformed individuals, and acquiring information about whether acts are subject to sanctions", Journal of Law, Economics, & Organization 6:93–128.
Kaplow, L. (1992a), "Government relief for risk associated with government action", Scandinavian Journal of Economics 94:525–541.
Kaplow, L. (1992b), "The optimal probability and magnitude of fines for acts that definitely are undesirable", International Review of Law and Economics 12:3–11.
Kaplow, L. (1992c), "Rules versus standards: an economic analysis", Duke Law Journal 42:557–629.
Kaplow, L. (1993), "Shifting plaintiffs' fees versus increasing damage awards", Rand Journal of Economics 24:625–630.
Kaplow, L. (1994a), "The value of accuracy in adjudication: an economic analysis", Journal of Legal Studies 23:307–401.
Kaplow, L. (1994b), "Optimal insurance contracts when establishing the amount of loss is costly", Geneva Papers on Risk and Insurance Theory 19:139–152.
Kaplow, L. (1995a), "A model of the optimal complexity of legal rules", Journal of Law, Economics, & Organization 11:150–163.
Kaplow, L. (1995b), "A note on subsidizing gifts", Journal of Public Economics 58:469–477.
Kaplow, L. (1996), "How tax complexity and enforcement affect the equity and efficiency of the income tax", National Tax Journal 49:135–150.
Kaplow, L. (2000), "General characteristics of rules", in: B. Bouckaert and G. De Geest, eds., Encyclopedia of Law and Economics, Vol. V (Edward Elgar, Cheltenham, UK) pp. 502–528.

Kaplow, L., and S. Shavell (1989), "Legal advice about information to present in litigation: its effects and social desirability", Harvard Law Review 102:565–615.
Kaplow, L., and S. Shavell (1990), "Legal advice about acts already committed", International Review of Law and Economics 10:149–159.
Kaplow, L., and S. Shavell (1992), "Private versus socially optimal provision of ex ante legal advice", Journal of Law, Economics, & Organization 8:306–320.
Kaplow, L., and S. Shavell (1994a), "Accuracy in the determination of liability", Journal of Law and Economics 37:1–15.
Kaplow, L., and S. Shavell (1994b), "Optimal law enforcement with self-reporting of behavior", Journal of Political Economy 102:583–606.
Kaplow, L., and S. Shavell (1994c), "Why the legal system is less efficient than the income tax in redistributing income", Journal of Legal Studies 23:667–681.
Kaplow, L., and S. Shavell (1996a), "Property rules versus liability rules: an economic analysis", Harvard Law Review 109:713–790.
Kaplow, L., and S. Shavell (1996b), "Accuracy in the assessment of damages", Journal of Law and Economics 39:191–210.
Kaplow, L., and S. Shavell (2000), "Should legal rules favor the poor? Clarifying the role of legal rules and the income tax in redistributing income", Journal of Legal Studies 29:821–835.
Kaplow, L., and S. Shavell (2001), "Any non-welfarist method of policy assessment violates the Pareto principle", Journal of Political Economy 109:281–286.
Kaplow, L., and S. Shavell (2002a), Fairness versus Welfare (Harvard University Press, Cambridge, MA); also in Harvard Law Review 114:961–1388.
Kaplow, L., and S. Shavell (2002b), "On the superiority of corrective taxes to quantity regulation", American Law and Economics Review 4, forthcoming.
Karpoff, J.M., and J.R. Lott (1999), "On the determinants and importance of punitive damage awards", Journal of Law and Economics 42:527–573.
Katz, A. (1987), "Measuring the demand for litigation: is the English rule really cheaper?" Journal of Law, Economics, & Organization 3:143–176.
Katz, A. (1988), "Judicial decisionmaking and litigation expenditure", International Review of Law and Economics 8:127–143.
Katz, A. (1990a), "The effect of frivolous lawsuits on the settlement of litigation", International Review of Law and Economics 10:3–27.
Katz, A. (1990b), "The strategic structure of offer and acceptance: game theory and the law of contract formation", Michigan Law Review 89:215–295.
Katz, A. (1990c), "Your terms or mine? The duty to read the fine print in contracts", Rand Journal of Economics 21:518–537.
Katz, A. (1993), "Transaction costs and the legal mechanics of exchange: when should silence in the face of an offer be construed as acceptance?" Journal of Law, Economics, & Organization 9:77–97.
Katz, A. (1996), "When should an offer stick? The economics of promissory estoppel in preliminary negotiations", Yale Law Journal 105:1249–1309.
Keeton, W.R., and E. Kwerel (1984), "Externalities in automobile insurance and the underinsured driver problem", Journal of Law and Economics 27:149–179.
Kennan, J., and R. Wilson (1993), "Bargaining with private information", Journal of Economic Literature 31:45–104.
Kessler, D., and S.D. Levitt (1999), "Using sentence enhancements to distinguish between deterrence and incapacitation", Journal of Law and Economics 42:343–363.
Kessler, D., and M. McClellan (1996), "Do doctors practice defensive medicine?" Quarterly Journal of Economics 111:353–390.
Khanna, V.S. (1996), "Corporate criminal liability: what purpose does it serve?" Harvard Law Review 109:1477–1534.

Kimenyi, M.S., W.F. Shughart and R.D. Tollison (1993), "What do judges maximize?" in: C.K. Rowley, ed., Public Choice Theory, Vol. 3, The Separation of Powers and Constitutional Political Economy (Elgar, Aldershot, UK) pp. 139–146.
Kitch, E.W. (1977), "The nature and function of the patent system", Journal of Law and Economics 20:265–290.
Klein, B., and K.B. Leffler (1981), "The role of market forces in assuring contractual performance", Journal of Political Economy 89:615–641.
Klein, B., R.G. Crawford and A.A. Alchian (1978), "Vertical integration, appropriable rents, and the competitive contracting process", Journal of Law and Economics 21:297–326.
Klemperer, P. (1990), "How broad should the scope of patent protection be?" Rand Journal of Economics 21:113–130.
Klerman, D. (1996), "Settling multidefendant lawsuits: the advantage of conditional setoff rules", Journal of Legal Studies 25:445–462.
Klevorick, A.K., M. Rothschild and C. Winship (1984), "Information processing and jury decisionmaking", Journal of Public Economics 23:245–278.
Klitgaard, R.E. (1988), Controlling Corruption (University of California Press, Berkeley).
Kobayashi, B.H., and J.R. Lott (1996), "In defense of criminal defense expenditures and plea bargaining", International Review of Law and Economics 16:397–416.
Kofman, F., and J. Lawarree (1993), "Collusion in hierarchical agency", Econometrica 61:629–656.
Kolstad, C.D., T. Ulen and G.V. Johnson (1990), "Ex post liability for harm vs. ex ante safety regulation: substitutes or complements?" American Economic Review 80:888–901.
Kornhauser, L.A. (1982), "An economic analysis of the choice between enterprise and personal liability for accidents", California Law Review 70:1345–1392.
Kornhauser, L.A. (1983), "Reliance, reputation, and breach of contract", Journal of Law and Economics 26:691–706.
Kornhauser, L.A. (1992a), "Modeling collegial courts I: path-dependence", International Review of Law and Economics 12:169–185.
Kornhauser, L.A. (1992b), "Modeling collegial courts II: legal doctrine", Journal of Law, Economics, & Organization 8:441–470.
Kornhauser, L.A., and R.L. Revesz (1994), "Multidefendant settlements: the impact of joint and several liability", Journal of Legal Studies 23:41–76.
Kraakman, R.H. (1984), "Corporate liability strategies and the costs of legal controls", Yale Law Journal 93:857–898.
Kraakman, R.H. (1986), "Gatekeepers: the anatomy of a third-party enforcement strategy", Journal of Law, Economics, & Organization 2:53–104.
Kremer, M. (1998), "Patent buyouts: a mechanism for encouraging innovation", Quarterly Journal of Economics 113:1137–1167.
Kronman, A.T. (1978a), "Mistake, disclosure, information, and the law of contracts", Journal of Legal Studies 7:1–34.
Kronman, A.T. (1978b), "Specific performance", University of Chicago Law Review 45:351–382.
Landes, E.M. (1982), "Insurance, liability, and accidents: a theoretical and empirical investigation of the effect of no-fault accidents", Journal of Law and Economics 25:49–65.
Landes, W.M. (1971), "An economic analysis of the courts", Journal of Law and Economics 14:61–107.
Landes, W.M. (1993), "Sequential versus unitary trials: an economic analysis", Journal of Legal Studies 22:99–134.
Landes, W.M., and R.A. Posner (1975), "The private enforcement of law", Journal of Legal Studies 4:1–46.
Landes, W.M., and R.A. Posner (1976), "Legal precedent: a theoretical and empirical analysis", Journal of Law and Economics 19:249–307.
Landes, W.M., and R.A. Posner (1978), "Salvors, finders, good samaritans, and other rescuers: an economic study of law and altruism", Journal of Legal Studies 7:83–128.

Landes, W.M., and R.A. Posner (1979), "Adjudication as a private good", Journal of Legal Studies 8:235–284.
Landes, W.M., and R.A. Posner (1981), "An economic theory of intentional torts", International Review of Law and Economics 1:127–154.
Landes, W.M., and R.A. Posner (1987a), The Economic Structure of Tort Law (Harvard University Press, Cambridge, MA).
Landes, W.M., and R.A. Posner (1987b), "Trademark law: an economic perspective", Journal of Law and Economics 30:265–309.
Landes, W.M., and R.A. Posner (1989), "An economic analysis of copyright law", Journal of Legal Studies 18:325–363.
Landsberger, M., and I. Meilijson (1982), "Incentive generating state dependent penalty system: the case of income tax evasion", Journal of Public Economics 19:333–352.
Langbein, J.H. (1985), "The German advantage in civil procedure", University of Chicago Law Review 52:823–866.
Leland, H.E. (1992), "Insider trading: should it be prohibited?" Journal of Political Economy 100: 859–887.
Levitt, S.D. (1996), "The effect of prison population size on crime rates: evidence from prison overcrowding litigation", Quarterly Journal of Economics 111:319–351.
Levitt, S.D. (1997a), "Using electoral cycles in police hiring to estimate the effect of police on crime", American Economic Review 87:270–290.
Levitt, S.D. (1997b), "Incentive compatibility constraints as an explanation for the use of prison sentences instead of fines", International Review of Law and Economics 17:179–192.
Levitt, S.D. (1998a), "Juvenile crime and punishment", Journal of Political Economy 106:1156–1185.
Levitt, S.D. (1998b), "Why do increased arrest rates appear to reduce crime: deterrence, incapacitation, or measurement error?" Economic Inquiry 36:353–372.
Lewis, T., and M. Poitevin (1997), "Disclosure of information in regulatory proceedings", Journal of Law, Economics, & Organization 13:50–73.
Libecap, G.D. (1986), "Property rights in economic history: implications for research", Explorations in Economic History 23:227–252.
Libecap, G.D. (1998), "Common property", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 1 (Macmillan, London) pp. 317–324.
Loewenstein, G., S. Issacharoff, C. Camerer and L. Babcock (1993), "Self-serving assessments of fairness and pretrial bargaining", Journal of Legal Studies 22:135–159.
Lott, J.R. (1992), "Do we punish high income criminals too heavily?" Economic Inquiry 30:583–608.
Lott, J.R., and D.B. Mustard (1997), "Crime, deterrence, and right-to-carry concealed handguns", Journal of Legal Studies 26:1–68.
Lueck, D. (1998), "First possession", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 2 (Macmillan, London) pp. 132–144.
Macey, J.R., and G.P. Miller (1991), "The plaintiffs' attorney's role in class action and derivative litigation: economic analysis and recommendations for reform", University of Chicago Law Review 58:1–118.
Machlup, F. (1958), "An economic review of the patent system", Study of the subcommittee on patents, trademarks, and copyrights, Committee on the Judiciary, United States Senate, study no. 15 (United States Government Printing Office, Washington, D.C.).
MacLeod, W.B., and J.M. Malcomson (1993), "Investments, holdup, and the form of market contracts", American Economic Review 83:811–837.
Magat, W.A., and W.K. Viscusi (1992), Informational Approaches to Regulation (MIT Press, Cambridge, MA).
Malik, A.S. (1990), "Avoidance, screening and optimum enforcement", Rand Journal of Economics 21:341–353.
Malik, A.S. (1993), "Self-reporting and the design of policies for regulating stochastic pollution", Journal of Environmental Economics and Management 24:241–257.

Maskin, E., and J. Moore (1999), "Implementation and renegotiation", Review of Economic Studies 66:39–56.
Mathios, A.D. (2000), "The impact of mandatory disclosure laws on product choices: an analysis of the salad dressing market", Journal of Law and Economics 43:651–677.
McMillan, J. (1994), "Selling spectrum rights", Journal of Economic Perspectives 8(3):145–162.
Menell, P. (1983), "A note on private versus social incentives to sue in a costly legal system", Journal of Legal Studies 12:41–52.
Menell, P. (2000), "Intellectual property: general theories", in: B. Bouckaert and G. De Geest, eds., Encyclopedia of Law and Economics, Vol. II (Edward Elgar, Cheltenham, UK) pp. 129–188.
Meurer, M.J. (1992), "The gains from faith in an unfaithful agent: settlement conflicts between defendants and liability insurers", Journal of Law, Economics, & Organization 8:502–522.
Miceli, T.J. (1993), "Optimal deterrence of nuisance suits by repeat defendants", International Review of Law and Economics 13:135–144.
Miceli, T.J. (1996), "Plea bargaining and deterrence: an institutional approach", European Journal of Law and Economics 3:249–264.
Miceli, T.J. (1997), Economics of the Law: Torts, Contracts, Property, Litigation (Oxford University Press, New York).
Miceli, T.J., and K. Segerson (1996), Compensation for Regulatory Takings: An Economic Analysis with Applications (JAI Press, Greenwich, CT).
Milgrom, P.R. (1981), "Good news and bad news: representation theorems and applications", Bell Journal of Economics 12:380–391.
Milgrom, P.R., and J. Roberts (1986), "Relying on the information of interested parties", Rand Journal of Economics 17:18–32.
Miller, G.P. (1986), "An economic analysis of rule 68", Journal of Legal Studies 15:93–125.
Miller, G.P. (1987), "Some agency problems in settlement", Journal of Legal Studies 16:189–215.
Miller, G.P. (1998), "Class actions", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 1 (Macmillan, London) pp. 257–262.
Mnookin, R.H. (1993), "Why negotiations fail: an exploration of barriers to the resolution of conflict", Ohio State Journal of Dispute Resolution 8:235–256.
Mnookin, R.H., and R. Wilson (1998), "A model of efficient discovery", Games and Economic Behavior 25:219–250.
Montesquieu, C. (1748), The Spirit of the Laws, 1977 reprint edition (University of California Press, Berkeley).
Mookherjee, D. (1997), "The economics of enforcement", in: A. Bose, M. Rakshit, and A. Sinha, eds., Issues in Economic Theory and Public Policy: Essays in Honor of Professor Tapas Majumdar (Oxford University Press, Delhi) pp. 202–249.
Mookherjee, D., and I.P.L. Png (1992), "Monitoring vis-à-vis investigation in enforcement of law", American Economic Review 82:556–565.
Mookherjee, D., and I.P.L. Png (1994), "Marginal deterrence in enforcement of law", Journal of Political Economy 102:1039–1066.
Mookherjee, D., and I.P.L. Png (1995), "Corruptible law enforcers: how should they be compensated?" Economic Journal 105:145–159.
Moore, M.J., and W.K. Viscusi (1990), Compensation Mechanisms for Job Risks: Wages, Workers' Compensation, and Product Liability (Princeton University Press, Princeton).
Murphy, R.S., and E.A. O'Hara (1997), "Mistake of federal criminal law: a study of coalitions and costly information", Supreme Court Economic Review 5:217–278.
Nagin, D. (1978), "General deterrence: a review of the empirical evidence", in: A. Blumstein, J. Cohen and D. Nagin, eds., Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates (National Academy of Sciences, Washington, D.C.) pp. 95–139.
Netter, J.M. (1998), "Adverse possession", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 1 (Macmillan, London) pp. 18–21.

Newman, H.A., and D.W. Wright (1990), "Strict liability in a principal–agent model", International Review of Law and Economics 10:219–231.
Newman, P., ed. (1998), New Palgrave Dictionary of Economics and the Law, Vols. 1–3 (Macmillan, London).
Noldeke, G., and K.M. Schmidt (1995), "Option contracts and renegotiation: a solution to the holdup problem", Rand Journal of Economics 26:163–179.
Ordover, J.A. (1978), "Costly litigation in the model of single activity accidents", Journal of Legal Studies 7:243–261.
Osborne, E. (1999), "Who should be worried about asymmetric information in litigation?" International Review of Law and Economics 19:399–409.
Ostrom, B.J., and N.B. Kauder (1996), Examining the Work of State Courts 1994 (National Center for State Courts, Williamsburg, VA).
Ostrom, E. (1998), "Self-governance of common-pool resources", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 3 (Macmillan, London) pp. 424–433.
Pigou, A.C. (1932), The Economics of Welfare (Macmillan, New York).
Pitchford, R. (1995), "How liable should a lender be? The case of judgment-proof firms and environmental risk", American Economic Review 85:1171–1186.
Png, I.P.L. (1986), "Optimal subsidies and damages in the presence of judicial error", International Review of Law and Economics 6:101–105.
Polinsky, A.M. (1980a), "Private versus public enforcement of fines", Journal of Legal Studies 9:105–127.
Polinsky, A.M. (1980b), "On the choice between property rules and liability rules", Economic Inquiry 18:233–246.
Polinsky, A.M. (1983), "Risk sharing through breach of contract remedies", Journal of Legal Studies 12:427–444.
Polinsky, A.M. (1989), An Introduction to Law and Economics, 2nd edition (Little, Brown and Company, Boston).
Polinsky, A.M. (1997), "Are punitive damages really insignificant, predictable, and rational?" Journal of Legal Studies 26:663–677.
Polinsky, A.M., and Y.-K. Che (1991), "Decoupling liability: optimal incentives for care and litigation", Rand Journal of Economics 22:562–570.
Polinsky, A.M., and D.L. Rubinfeld (1988a), "The welfare implications of costly litigation for the level of liability", Journal of Legal Studies, 17:151–164.
Polinsky, A.M., and D.L. Rubinfeld (1988b), "The deterrent effects of settlements and trials", International Review of Law and Economics 8:109–116.
Polinsky, A.M., and D.L. Rubinfeld (1991), "A model of optimal fines for repeat offenders", Journal of Public Economics 46:291–306.
Polinsky, A.M., and D.L. Rubinfeld (1993), "Sanctioning frivolous suits: an economic analysis", Georgetown Law Journal 82:397–435.
Polinsky, A.M., and D.L. Rubinfeld (1996), "Optimal awards and penalties when the probability of prevailing varies among plaintiffs", Rand Journal of Economics 27:269–280.
Polinsky, A.M., and S. Shavell (1979), "The optimal tradeoff between the probability and magnitude of fines", American Economic Review 69:880–891.
Polinsky, A.M., and S. Shavell (1981), "Contribution and claim reduction among antitrust defendants: an economic analysis", Stanford Law Review 33:447–471.
Polinsky, A.M., and S. Shavell (1984), "The optimal use of fines and imprisonment", Journal of Public Economics 24:89–99.
Polinsky, A.M., and S. Shavell (1992), "Enforcement costs and the optimal magnitude and probability of fines", Journal of Law and Economics 35:133–148.
Polinsky, A.M., and S. Shavell (1993), "Should employees be subject to fines and imprisonment given existence of corporate liability?" International Review of Law and Economics 13:239–257.

Polinsky, A.M., and S. Shavell (1998a), "Punitive damages: an economic analysis", Harvard Law Review 111:869–962.
Polinsky, A.M., and S. Shavell (1998b), "On offense history and the theory of deterrence", International Review of Law and Economics 18:305–324.
Polinsky, A.M., and S. Shavell (1999), "On the disutility and discounting of imprisonment and the theory of deterrence", Journal of Legal Studies 28:1–16.
Polinsky, A.M., and S. Shavell (2000), "The economic theory of public enforcement of law", Journal of Economic Literature 38:45–76.
Polinsky, A.M., and S. Shavell (2001), "Corruption and optimal law enforcement", Journal of Public Economics 81:1–24.
Posner, E.A. (1995), "Contract law in the welfare state: a defense of the unconscionability doctrine, usury laws, and related limitations on the freedom to contract", Journal of Legal Studies 24:283–319.
Posner, E.A. (1997), "Altruism, status, and trust in the law of gifts and gratuitous promises", Wisconsin Law Review 1997:567–609.
Posner, R.A. (1972), Economic Analysis of Law, 1st edition (Little, Brown and Company, Boston).
Posner, R.A. (1973), "An economic approach to legal procedure and judicial administration", Journal of Legal Studies 2:399–458.
Posner, R.A. (1975), "The economic approach to law", Texas Law Review 53:757–782.
Posner, R.A. (1977a), Economic Analysis of Law, 2nd edition (Little, Brown and Company, Boston).
Posner, R.A. (1977b), "Gratuitous promises in economics and law", Journal of Legal Studies 6:411–426.
Posner, R.A. (1985), "An economic theory of the criminal law", Columbia Law Review 85:1193–1231.
Posner, R.A. (1986), "The summary jury trial and other methods of alternative dispute resolution: some cautionary observations", University of Chicago Law Review 53:366–393.
Posner, R.A. (1993a), "What do judges and justices maximize? (The same thing everybody else does)", Supreme Court Economic Review 3:1–41.
Posner, R.A. (1993b), "Blackmail, privacy, and freedom of contract", University of Pennsylvania Law Review 141:1817–1847.
Posner, R.A. (1998), Economic Analysis of Law, 5th edition (Aspen Law and Business, New York).
Posner, R.A. (1999), "An economic approach to the law of evidence", Stanford Law Review 51: 1477–1546.
Posner, R.A., and A.M. Rosenfield (1977), "Impossibility and related doctrines in contract law: an economic analysis", Journal of Legal Studies 6:83–118.
Pound, R. (1959), An Introduction to the Philosophy of Law (Yale University Press, New Haven).
Priest, G.L. (1977), "The common law process and the selection of efficient rules", Journal of Legal Studies 6:65–82.
Priest, G.L. (1981), "A theory of the consumer product warranty", Yale Law Journal 90:1297–1352.
Priest, G.L. (1988), "Products liability law and the accident rate", in: R.E. Litan and C. Winston, eds., Liability: Perspectives and Policy (Brookings Institution, Washington, D.C.) pp. 184–222.
Priest, G.L., and B. Klein (1984), "The selection of disputes for litigation", Journal of Legal Studies 13:1–55.
Pyle, D.J. (1983), The Economics of Crime and Law Enforcement (Macmillan, London).
Rabin, M. (1998), "Psychology and economics", Journal of Economic Literature 36:11–46.
Raiffa, H. (1968), Decision Analysis (Addison-Wesley, Reading, MA).
Ramseyer, J.M. (1998), "Judicial independence", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 2 (Macmillan, London) pp. 383–387.
Ramseyer, J.M., and M. Nakazato (1989), "The rational litigant: settlement amounts and verdict rates in Japan", Journal of Legal Studies 18:263–290.
Rasmusen, E. (1994), "Judicial legitimacy as a repeated game", Journal of Law, Economics, & Organization 10:63–83.
Reinganum, J.F. (1988), "Plea bargaining and prosecutorial discretion", American Economic Review 78:713–728.

Reinganum, J.F. (1989), "The timing of innovation: research, development, and diffusion", in: R. Schmalensee and R.D. Willig, eds., Handbook of Industrial Organization, Vol. 1 (North-Holland, Amsterdam) pp. 849–908.
Reinganum, J.F. (2000), "Sentencing guidelines, judicial discretion, and plea bargaining", Rand Journal of Economics 31:62–81.
Reinganum, J.F., and L.L. Wilde (1986), "Settlement, litigation, and the allocation of litigation costs", Rand Journal of Economics 17:557–566.
Roberts, M.J., and M. Spence (1976), "Effluent charges and licenses under uncertainty", Journal of Public Economics 5:193–208.
Rogerson, W.P. (1984), "Efficient reliance and damage measures for breach of contract", Rand Journal of Economics 15:39–53.
Rogerson, W.P. (1992), "Contractual solutions to the hold-up problem", Review of Economic Studies 59:777–794.
Rose-Ackerman, S. (1978), Corruption: A Study in Political Economy (Academic Press, New York).
Rose-Ackerman, S. (1985), Inalienability and the theory of property rights", Columbia Law Review 85:931–969.
Rose-Ackerman, S. (1999), Corruption and Government: Causes, Consequences and Reform (Cambridge University Press, New York).
Rose-Ackerman, S., and M. Geistfeld (1987), "The divergence between social and private incentives to sue: a comment on Shavell, Menell and Kaplow", Journal of Legal Studies 16:483–491.
Rosenberg, D. (1984), "The causal connection in mass exposure cases: a 'public law' vision of the tort system", Harvard Law Review 97:849–929.
Rosenberg, D., and S. Shavell (1985), "A model in which suits are brought for their nuisance value", International Review of Law and Economics 5:3–13.
Rubin, P.H. (1977), "Why is the common law efficient?" Journal of Legal Studies 6:51–63.
Rubin, P.H. (1993), Tort Reform by Contract (AEI Press, Washington, D.C.).
Rubinfeld, D.L., and D.E.M. Sappington (1987), "Efficient awards and standards of proof in judicial proceedings", Rand Journal of Economics 18:308–315.
Rubinfeld, D.L., and S. Scotchmer (1993), "Contingent fees for attorneys: an economic analysis", Rand Journal of Economics 24:343–356.
Rubinfeld, D.L., and S. Scotchmer (1998), "Contingent fees", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 1 (Macmillan, London) pp. 415–420.
Rubinstein, A. (1979), "An optimal conviction policy for offenses that may have been committed by accident", in: S.J. Brams, A. Schotter and G. Schwodiauer, eds., Applied Game Theory (Physica-Verlag, Wurzburg) pp. 406–413.
Sah, R.K. (1991), "Social osmosis and patterns of crime", Journal of Political Economy 99:1272–1295.
Sanchirico, C.W. (2000), "Taxes versus legal rules as instruments for equity: a more equitable view", Journal of Legal Studies 29:797–820.
Scherer, F.M., and D. Ross (1990), Industrial Market Structure and Economic Performance, 3rd edition (Houghton Mifflin, Boston).
Schmitz, P.W. (2000), "On the joint use of liability and safety regulation", International Review of Law and Economics 20:371–382.
Schrag, J.L. (1999), "Managerial judges: an economic analysis of the judicial management of legal discovery", Rand Journal of Economics 30:305–323.
Schrag, J.L., and S. Scotchmer (1994), "Crime and prejudice: the use of character evidence in criminal trials", Journal of Law, Economics, & Organization 10:319–342.
Schwartz, A. (1979), "The case for specific performance", Yale Law Journal 89:271–306.
Schwartz, A. (1992), "Relational contracts in the courts: an analysis of incomplete agreements and judicial strategies", Journal of Legal Studies 21:271–318.
Schwartz, A., and L.L. Wilde (1979), "Intervening in markets on the basis of imperfect information: a legal and economic analysis", University of Pennsylvania Law Review 127:630–682.

Schwartz, E.P., and W.F. Schwartz (1996), "The challenge of peremptory challenges", Journal of Law, Economics, & Organization 12:325–360.
Schweizer, U. (1989), "Litigation and settlement under two-sided incomplete information", Review of Economic Studies 56:163–178.
Scotchmer, S. (1996), "Protecting early innovators: should second-generation products be patentable?" Rand Journal of Economics 27:322–331.
Scotchmer, S. (1999), "On the optimality of the patent renewal system", Rand Journal of Economics 30:181–196.
Scott, K.E. (1998), "Insider trading", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law, Vol. 2 (Macmillan, London) pp. 326–330.
Segal, I., and M.D. Whinston (2001), "The Mirrlees approach to mechanism design with renegotiation (with applications to hold-up and risk sharing)", Econometrica, forthcoming.
Sen, A., and B. Williams (1982), Utilitarianism and Beyond (Cambridge University Press, Cambridge, UK).
Sethi, R., and E. Somanathan (1996), "The evolution of social norms in common property resource use", American Economic Review 86:766–788.
Shavell, S. (1980a), "An analysis of causation and the scope of liability in the law of torts", Journal of Legal Studies 9:463–516.
Shavell, S. (1980b), "Damage measures for breach of contract", Bell Journal of Economics 11:466–490.
Shavell, S. (1980c), "Strict liability versus negligence", Journal of Legal Studies 9:1–25.
Shavell, S. (1981), "A note on efficiency vs. distributional equity in legal rulemaking: should distributional equity matter given optimal income taxation?" American Economic Association Papers and Proceedings 71:414–418.
Shavell, S. (1982a), "On liability and insurance", Bell Journal of Economics 13:120–132.
Shavell, S. (1982b), "The social versus the private incentive to bring suit in a costly legal system", Journal of Legal Studies 11:333–339.
Shavell, S. (1982c), "Suit, settlement, and trial: a theoretical analysis under alternative methods for the allocation of legal costs", Journal of Legal Studies 11:55–81.
Shavell, S. (1984a), "A model of the optimal use of liability and safety regulation", Rand Journal of Economics 15:271–280.
Shavell, S. (1984b), "The design of contracts and remedies for breach", Quarterly Journal of Economics 99:121–148.
Shavell, S. (1984c), "Liability for harm versus regulation of safety", Journal of Legal Studies 13:357–374.
Shavell, S. (1985a), "Criminal law and the optimal use of nonmonetary sanctions as a deterrent", Columbia Law Review 85:1232–1262.
Shavell, S. (1985b), "Uncertainty over causation and the determination of civil liability", Journal of Law and Economics 28:587–609.
Shavell, S. (1986), "The judgment proof problem", International Review of Law and Economics 6:45–58.
Shavell, S. (1987a), Economic Analysis of Accident Law (Harvard University Press, Cambridge, MA).
Shavell, S. (1987b), "The optimal use of nonmonetary sanctions as a deterrent", American Economic Review 77:584–592.
Shavell, S. (1987c), "A model of optimal incapacitation", American Economic Review 77:107–110.
Shavell, S. (1988), "Legal advice about contemplated acts: the decision to obtain advice, its social desirability, and protection of confidentiality", Journal of Legal Studies 17:123–150.
Shavell, S. (1989a), "Sharing of information prior to settlement or litigation", Rand Journal of Economics 20:183–195.
Shavell, S. (1989b), "Optimal sanctions and the incentive to provide evidence to legal tribunals", International Review of Law and Economics 9:3–11.
Shavell, S. (1990), "Deterrence and the punishment of attempts", Journal of Legal Studies 19:435–466.
Shavell, S. (1991a), "An economic analysis of altruism and deferred gifts", Journal of Legal Studies 20:401–421.

Shavell, S. (1991b), "Specific versus general enforcement of law", Journal of Political Economy 99:1088–1108.
Shavell, S. (1991c), "Individual precautions to prevent theft: private versus socially optimal behavior", International Review of Law and Economics 11:123–132.
Shavell, S. (1992), "A note on marginal deterrence", International Review of Law and Economics 12:345–355.
Shavell, S. (1993a), "The optimal structure of law enforcement", Journal of Law and Economics 36:255–287.
Shavell, S. (1993b), "Suit versus settlement when parties seek nonmonetary judgments", Journal of Legal Studies 22:1–13.
Shavell, S. (1993c), "An economic analysis of threats and their illegality: blackmail, extortion, and robbery", University of Pennsylvania Law Review 141:1877–1903.
Shavell, S. (1994), "Acquisition and disclosure of information prior to sale", Rand Journal of Economics 25:20–36.
Shavell, S. (1995a), "Alternative dispute resolution: an economic analysis", Journal of Legal Studies 24:1–28.
Shavell, S. (1995b), "The appeals process as a means of error correction", Journal of Legal Studies 24:379–426.
Shavell, S. (1996), "Any frequency of plaintiff victory at trial is possible", Journal of Legal Studies 25:493–501.
Shavell, S. (1997), "The fundamental divergence between the private and the social motive to use the legal system", Journal of Legal Studies 26:575–612.
Shavell, S. (1999), "The level of litigation: private versus social optimality", International Review of Law and Economics 19:99–115.
Shavell, S. (2000), "On the social function and the regulation of liability insurance", Geneva Papers on Risk and Insurance, Issues and Practice 25:166–179.
Shavell, S. (2002), Principles of Economic Analysis of Law (Harvard University Press, Cambridge, MA) forthcoming.
Shavell, S., and T. van Ypersele (2001), "Rewards versus intellectual property rights", Journal of Law and Economics 44:525–547.
Shepherd, G.B. (1999), "An empirical study of the economics of pretrial discovery", International Review of Law and Economics 19:245–263.
Shin, H.S. (1998), "Adversarial and inquisitorial procedures in arbitration", Rand Journal of Economics 29:378–405.
Shleifer, A. (1998), "State versus private ownership", Journal of Economic Perspectives 12(4):133–150.
Shleifer, A., and R.W. Vishny (1993), "Corruption", Quarterly Journal of Economics 108:599–617.
Sieg, H. (2000), "Estimating a bargaining model with asymmetric information: evidence from medical malpractice disputes", Journal of Political Economy 108:1006–1021.
Silver, C. (2000), "Class actions-representative proceedings", in: B. Bouckaert and G. De Geest, eds., Encyclopedia of Law and Economics, Vol. V (Edward Elgar, Cheltenham, UK) pp. 194–240.
Sloan, F.A., B.A. Reilly and C.M. Schenzler (1994), "Tort liability versus other approaches for deterring careless driving", International Review of Law and Economics 14:53–71.
Sobel, J. (1985), "Disclosure of evidence and resolution of disputes: who should bear the burden of proof?" in: A.E. Roth, ed., Game-Theoretic Models of Bargaining (Cambridge University Press, Cambridge, UK) pp. 341–361.
Sobel, J. (1989), "An analysis of discovery rules", Law and Contemporary Problems 52:133–159.
Spence, M. (1977), "Consumer misperceptions, product failure, and product liability", Review of Economic Studies 44:561–572.
Spier, K.E. (1992a), "The dynamics of pretrial negotiation", Review of Economic Studies 59:93–108.
Spier, K.E. (1992b), "Incomplete contracts and signalling", Rand Journal of Economics 23:432–443.

Spier, K.E. (1994a), "Pretrial bargaining and the design of fee-shifting rules", Rand Journal of Economics 25:197–214.
Spier, K.E. (1994b), "Settlement bargaining and the design of damage awards", Journal of Law, Economics, & Organization 10:84–95.
Spier, K.E. (1997), "A note on the divergence between the private and the social motive to settle under a negligence rule", Journal of Legal Studies 26:613–621.
Spier, K.E., and M.D. Whinston (1995), "On the efficiency of privately stipulated damages for breach of contract: entry barriers, reliance, and renegotiation", Rand Journal of Economics 26:180–202.
Spitzer, M., and E. Talley (2000), "Judicial auditing", Journal of Legal Studies 29:649–683.
Stigler, G.J. (1970), "The optimum enforcement of laws", Journal of Political Economy 78:526–536.
Stole, L.A. (1992), "The economics of liquidated damages clauses in contractual environments with private information", Journal of Law, Economics, & Organization 8:582–606.
Sykes, A.O. (1981), "An efficiency analysis of vicarious liability under the law of agency", Yale Law Journal 91:168–206.
Sykes, A.O. (1990), "The doctrine of commercial impracticability in a second-best world", Journal of Legal Studies 19:43–94.
Sykes, A.O. (1994), "'Bad faith' refusal to settle by liability insurers: some implications of the judgment-proof problem", Journal of Legal Studies 23:77–110.
Tauchen, H.V., A.D. Witte and H. Griesinger (1994), "Criminal deterrence: revisiting the issue with a birth cohort", Review of Economics and Statistics 76:399–412.
Tirole, J. (1986), "Hierarchies and bureaucracies: on the role of collusion in organizations", Journal of Law, Economics, & Organization 2:181–214.
Tirole, J. (1988), The Theory of Industrial Organization (MIT Press, Cambridge, MA).
Trebilcock, M.J. (1993), The Limits of Freedom of Contract (Harvard University Press, Cambridge, MA).
Ulen, T. (1984), "The efficiency of specific performance: toward a unified theory of contract remedies", Michigan Law Review 83:341–403.
Umbeck, J.R. (1981), A Theory of Property Rights with Application to the California Gold Rush (Iowa State University Press, Ames).
van Wijck, P., and B. van Velthoven (2000), "An economic analysis of the American and the continental rule for allocating legal costs", European Journal of Law and Economics 9:115–125.
Viscusi, W.K. (1986a), "The determinants of the disposition of product liability claims and compensation for bodily injury", Journal of Legal Studies 15:321–346.
Viscusi, W.K. (1986b), "The risks and rewards of criminal activity: a comprehensive test of criminal deterrence, Part I", Journal of Labor Economics 4:317–340.
Viscusi, W.K. (1988), "Product liability litigation with risk aversion", Journal of Legal Studies 17: 101–121.
Viscusi, W.K. (1991), Reforming Products Liability (Harvard University Press, Cambridge, MA).
Viscusi, W.K., and W. Evans (1990), "Utility functions that depend on health status: estimates and economic implications", American Economic Review 80:353–374.
Viscusi, W.K., and W.A. Magat (1987), Learning About Risk: Consumer and Worker Responses to Hazard Information (Harvard University Press, Cambridge, MA).
Viscusi, W.K., J.M. Vernon and J.E. Harrington Jr (1995), Economics of Regulation and Antitrust, 2nd edition (MIT Press, Cambridge, MA).
Waldfogel, J. (1995a), "Are fines and prison terms used efficiently? Evidence on federal fraud offenders", Journal of Law and Economics 38:107–139.
Waldfogel, J. (1995b), "The selection hypothesis and the relationship between trial and plaintiff victory", Journal of Political Economy 103:229–260.
Weitzman, M.L. (1974), "Prices vs. quantities", Review of Economic Studies 41:477–491.
Wilde, L.L. (1992), "Criminal choice, nonmonetary sanctions, and marginal deterrence: a normative analysis", International Review of Law and Economics 12:333–344.

Wilkins, D.B. (1992), "Who should regulate lawyers?" Harvard Law Review 105:799–887.
Williamson, O.E. (1975), Markets and Hierarchies, Analysis and Antitrust Implications: A Study in the Economics of Internal Organization (Free Press, New York).
Williamson, O.E. (1985), The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting (Free Press, New York).
Wils, W.P.J. (1993), "Who should bear the costs of failed negotiations? A functional inquiry into precontractual liability", Journal des Economistes et des Etudes Humaines 4:93–134.
Wilson, J.Q., and R.J. Herrnstein (1985), Crime and Human Nature (Simon and Schuster, New York).
Witte, A.D. (1980), "Estimating the economic model of crime with individual data", The Quarterly Journal of Economics 94:57–84.
Wittman, D. (1977), "Prior regulation versus post liability: the choice between input and output monitoring", Journal of Legal Studies 6:193–212.
Wittman, D. (1981), "Optimal pricing of sequential inputs: last clear chance, mitigation of damages, and related doctrines in the law", Journal of Legal Studies 10:65–91.
Wittman, D. (1985), "Is the selection of cases for trial biased?" Journal of Legal Studies 14:185–214.
Wright, B.D. (1983), "The economics of invention incentives: patents, prizes, and research contracts", American Economic Review 73:691–707.

AUTHOR INDEX

SUBJECT INDEX