Toward a Theory of Optimal Tax Systems

Joel Slemrod

June 17, 2008

To be presented at the 2008 Conference of the New Zealand Association of Economists and the Australasian Meeting of the Econometric Society, "Markets and Models: Policy Frontiers in the AWH Phillips Tradition," Wellington, New Zealand, July 9-11, 2008. I am grateful to Henrik Kleven for comments and discussion about an earlier draft.

I. Introduction

In a remarkable half decade in the first half of the 1970s, the modern normative theory of taxation, known as optimal taxation, was established. It placed the evaluation of taxation on a rigorous footing. To be sure, rigorous analysis certainly preceded this development, but much of the earlier analysis addressed the positive, or descriptive, side of taxation (i.e., analyzing the consequences of tax policies), and some rigorous normative analysis preceded it, such as Frank Ramsey's work in the 1920s, Corlett and Hague's work in the 1950s, and so on.

Underlying the modern theory were several key assumptions, which one may group into two sets as follows:

Set #1:

- Focus on benevolent government.
- People understand and react rationally to the tax system. Thus the government has no reason to manipulate citizens' perceptions.
- A consequentialist and welfarist orientation.

Set #2:

- The central role of information, and in particular asymmetric information between the government and private citizens.
- A sharp distinction between what is measurable without cost and what is not measurable at any cost. For example, optimal income tax models presume that the government is assumed to be able to observe

income without cost, but cannot at any cost observe hours, wage rates, or ability types.

- No meaningful role for firms.

Optimal taxation, pioneered by Peter Diamond, Nobel-laureate James Mirrlees and others, was an elegant theoretical structure that enabled economists to make rigorous statements about what tax system would best achieve explicit objectives under carefully laid out stylized conditions. Here are two examples, which I'll return to.

The first concerns production efficiency. Diamond and Mirrlees (1971) established that, although in the absence of lump-sum taxes there will certainly be distortions from raising revenue, under some conditions the second-best optimum will always feature production efficiency: whatever goods and services are consumed should be produced (or obtained, when imports are an option) in a social-cost-minimizing way. Second, in an optimal commodity tax framework, in general all taxed goods should be taxed differentially or, in other words, uniform taxation is optimal only under very strong assumptions about utility functions.

Perhaps most importantly, optimal taxation reasoning replaced vague argument that did not lend itself to intellectual progress. My favorite example of the imprecision of pre-optimal tax normative reasoning comes from a 1917 book by Robert Jones entitled *The Nature and First Principle of Taxation* in which the author offers a list (incomplete and overlapping, he admits) of seventeen "maxims, canons and principles" of taxation stated in the literature to that day. They are: equality, proportionality, certainty, economy, convenience, productivity, justice or equity, generality, consistence, elasticity, unity, diffusion, exemption of minimum, relation to

franchise, graduation, minimum sacrifice, and faculty or ability. No modern economist would have the patience to argue which tax system better achieved, or much less struck a balance, among these 17 principles. In the framework of the theory of optimal taxation, though, one can have a productive dialogue about what these principles mean, and how alternative tax systems trade them off, and how the tradeoff depend son clear assumptions about the model of the economy and its parameterization.

For all of its important contributions to our understanding, now—nearly four decades after the birth of optimal taxation theory--both sets of underlying assumptions are under attack. The public choice economists have argued convincingly that the individuals who comprise the government (i.e., bureaucrats and politicians) are self-motivated and rational—or at least as rational as everyone else—and therefore do not automatically make decisions in the public interest.

More recently, proponents of "behavioral" economics have questioned the core assumption about rational decision-making that underlies not only public economics, but all of economics. Although both state and non-state actors may have bounded rationality, there are undoubtedly aspects of behavioral *public* economics that are unique, such as the tendency of people to mindlessly follow authority, as the unsettling experiments of Milgrom showed half a century ago. It is also true that, consciously or not, tax systems have many features that could have been designed by the marketing director of Procter and Gamble, such as the use of discounts in calculating taxable income, the use of tax refunds to provide salience to receiving rather than remitting funds to the government, and the use of withholding to make tax liabilities feel more like an installment sale than a cash-on-delivery sale.

Today I will focus on the second set of assumptions underlying optimal taxation theory. I will address what I see as inadequate in the standard framework, and discuss some recent attempts to rebuild it. I will first provide an overview of my theme, and then discuss in more detail examples of a research agenda designed to reconstruct the theoretical and empirical wings of tax analysis.

To help explain my agenda, I will use an analogy to physics, and discuss what are known as correspondence principles. In physics, classical theories like classical mechanics and classical electrodynamics accurately describe macroscopic systems like springs and capacitors. However, in describing microscopic objects, such as atoms and elementary particles, the distinct rules of quantum mechanics are highly successful. The Danish physicist Niels Bohr argued that the laws of physics should be independent of the size of the physical objects being described, so there must be some limit in which quantum mechanics reduces to classical mechanics. Bohr's correspondence principle, which he formulated explicitly in 1923, demands that classical physics and quantum physics give the same answer when the systems reach a certain size.

In natural science the term "correspondence principle" is used in a more general sense to mean that a new scientific theory reduces to an earlier scientific theory in appropriate circumstances. This requires that the new theory explain all the phenomena under circumstances for which the preceding theory was known to be valid, the "correspondence limit."

I suggest that in tax theory two correspondence principles should apply. The first is that the theory should apply to both developed and

developing countries, where the two settings differ both in the cost of acquiring information about tax bases and in the administrative capacity of the tax authority. To be sure, there is a long, rich, and well-informed literature on the importance of administrative considerations in developing countries, some of which focuses on the role of so-called "tax handles," tax bases that are easily measurable and for which the liability is easily collectible, but are not otherwise ideal. A short list of tax phenomena that are common in developing countries because of administrative considerations are presumptive taxes, taxes collected at the border, and size cutoffs for inclusion in the business and personal tax net, explicit or not. But the rigorous normative theory has not ventured very far into this different setting.

We also need a correspondence principle that links the model of the behavioral response to taxation of sophisticated taxpayers such as high-income individuals and multinational corporations, to the model of the behavioral response of most everyone else. This is especially important after a quarter of a century of growing income inequality in many countries; in the U.S. now the top 1% of income recipients receive 21% of adjusted gross income, and owe 39% of federal income tax liability. We need a positive and normative model to address both the CEO and hedge fund manager that have access to, and appetite, for sophisticated tax avoidance techniques, as well as the wage earner whose income tax liability is remitted by his employer and the small business that can effectively stay under the radar of the tax authority.

One way to state what I am seeking to replace a theory of taxation with a theory of tax systems. I begin by defining the latter.

A Theory of Tax Systems

A tax system is a set of rules, regulations, and procedures that sets out three things:

1) It defines what events or states of the world trigger tax liability, and the magnitude of that liability: *tax bases and rates*.

2) It defines who or what entity must remit that tax liability, and when: *remittance rules*.

3) It defines the procedures that facilitate and ensure compliance with the remittance rules, including information reporting requirements and the consequences (including penalties) of not remitting the liability in a timely fashion: *administrative and enforcement rules*.

Most modern economic analysis of taxation presumes that tax liability can be ascertained and collected costlessly, in which case 2) is irrelevant and 3) is unnecessary. But, in reality, governments have limited administrative capacity to measure, monitor, and enforce and evasion and avoidance are ubiquitous and administrative and compliance costs are not trivial. This is especially true in developing countries, but is true in every tax system.

Indeed, no government can announce a tax system and then rely on taxpayers' sense of duty to remit what is owed. Some dutiful people will undoubtedly pay what they owe, but many others will not. Over time the ranks of the dutiful will shrink, as they see how they are being taken advantage of by the others. Thus, paying taxes must be made a legal responsibility of citizens, with penalties attendant on non-compliance, and procedures must be put in place for the tax authority to receive information. But even in the face of those penalties, substantial tax evasion exists -- and always has.

A theory of tax systems would have to address much more than the optimal tax base (income or consumption) and the optimal rates to apply to that base. It would have to address such things as what fraction of tax returns to audit, how to choose the audited returns, and what structure of penalties to apply to detected evasion. It would have to address whether to have consumers remit retail sales taxes or retailers do, whether employers remit labor income tax or employees do. It would have to address what compromises to the ideal (in the absence of avoidance and evasion) base, such as the taxation of capital gains upon realization rather than accrual, the taxation, or non-taxation of the imputed income from owner-occupied housing, the use of statutory depreciation schedules rather than the true decline in value of capital assets, and so on.

A theory of tax systems would revisit the issue of tax remittance—who writes the checks to cover the tax liability. Our textbooks assert that a uniform value-added tax (VAT) and a retail sales tax (RST) are really equivalent tax systems. But they are not, because the remittance system—who writes the checks to the government—are different, and that turns out to be a crucial difference, and in large part explains why the VAT is the world tax success story of the last half century, adopted by over 140 countries, and why no country levies a RST at a rate exceeding 10%.

Who remits tax may be—especially, but not only, in developing countries—an important aspect of implementing a tax system, in spite of standard textbook assertions that which side of a taxed market remits a given amount of tax liability is completely irrelevant for the consequences—incidence, allocation and efficiency—of taxation. Except that this is not true. In the presence of costly information acquisition, which leads to avoidance and evasion (i.e., *in all real tax systems*), the cost of

administration and enforcement varies depending on the identity of the remitter. This can occur for two distinct reasons. The first is that the total resource costs of administering a given effective tax structure may vary depending on the remittance system. For example, it may be less costly (considering both administrative and compliance costs) to monitor one employer's tax remittances as opposed to thousands of employees.

Second, the opportunities for avoidance and evasion and the technology of the enforcement mechanism may affect the incentive to demand and supply the taxed activity. If this effect is not symmetric across the identity of the remitter, then changing the remittance system will affect the equilibrium price and quantity of the taxed activity, and thus have incidence and allocation effects.

A theory of tax systems would have to satisfy the two correspondence principles I have mentioned.

II. Towards Correspondence Principles

Avoidance and evasion

The focus of optimal tax theory is the compensated elasticity of the tax base to changes in the tax rate applied to it. The larger is this elasticity, the greater is the marginal efficiency cost per dollar raised from increasing the tax rate. This occurs because the behavioral response breaks the link between the social benefit of taxing (the revenue gained) and the cost to the taxpayers of taxing.

If there are many possible tax bases the higher is this elasticity of response, the lower should the tax rate on that base be. If there is only one possible base, say an income tax, the greater is the efficiency cost per dollar

raised, the smaller should government be, both in how much public goods it provides, and in how much redistribution it effects.

In the standard model, the central behavioral elasticity is the labor supply elasticity. Note that this is the *only* elasticity in Mirrlees' seminal contribution to optimal income tax progressivity, because the only decision an individual makes is how much—if at all—to work, which determines, for a given tax schedule, their after-tax income and consumption of goods.

Let me pause here to note what, for lack of a better term, I will call an irony about the pure theory of taxation. The key determinant of these two most politically charged, so ethically imbued, of policies—tax progressivity and the size of government—depend critically on this most mundane of concepts, the compensated behavioral response of labor supply, which in turn depends on the elasticity of substitution between leisure and goods. Who would have thought that the answer to this central question of political philosophy, the size and role of government, would depend on the shape of indifference curves? Rousseau?

But it does. And it does because, as George Will, the American writer and pundit, recently wrote in a column concerning U.S. policy in Iraq: "there can be no moral duty to do what cannot be done." The elasticity increases the social cost of raising revenue through taxation, be it for redistribution or for financing public goods, because it diminishes the (social) benefit of raising funds relative to the (private) loss. In the limit, when the tax authority faces a negatively-sloped Laffer curve, the behavioral response is so large that any further tax rate hike imposes costs on the taxpayer but collects *no* revenue that can be used to improve the utility of others, either via transfer or the financing of public goods. More generally, the base elasticity limits the effectiveness of what can be done, so that George Will's

statement really should be modified to read: “there can be no moral duty to do what cannot be done in a cost-effective way.”

Once it’s put that way, it is immediately clear that the cost of taxing relative to the potential benefit is increased not just by taxpayers’ shifting from consuming goods to consumed leisure, but by a wide range of behavioral responses induced by taxation. When higher tax rates send taxpayers to the Business School library looking for tax loopholes, or to the Cayman Islands looking for undetectable credit card accounts, the social benefit of the revenue relative to private cost is lower.

Since an important article by Martin Feldstein published in 1995, tax economists have has widely accepted the notion that, with respect to for example an income tax, all of these responses (labor supply, avoidance, evasion, etc.) can be usefully summarized by the elasticity of taxable income (ETI) and that, under certain assumptions, this ETI is a sufficient statistic for the marginal efficiency cost of higher income tax rates. This is because, at the margin, a taxpayer is willing to sacrifice utility valued at one dollar in order to reduce tax liability by one dollar. This sacrifice could take many forms, such as additional risk bearing due to evasion, expending real resources to identify and execute avoidance schemes, or substitution to activities that are more lightly taxed but less rewarding. The key insight of the ETI literature is that we do not need to know whether the behavioral response—the leak in revenue—is due to evasion, due to avoidance, or due to substitution in order to evaluate the costs to society. All one needs to know is potential tax revenue (assuming no behavioral change) 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.

The change in focus from the elasticity of labor supply to the elasticity of taxable income is especially critical for the second correspondence principle. Because of their prominence in revenue collected, importance to the economy and in debates about distributive justice, the behavioral response of high-income individuals is of particular importance. For this group the largest behavioral response is almost certainly not the labor supply decision, the traditional focus of behavioral response research, but rather sophisticated tax planning strategies that eliminate or defer taxable income, or convert ordinary income into preferentially-taxed capital gains, or outright evasion.

Recent econometric evidence suggests that the responsiveness of taxable income to tax rates, the elasticity of taxable income, is higher among the rich than among any other income group. This suggests that, other things equal, the marginal tax rate applied to these groups should be relatively low. Note we are now dangerously close to a real-live policy question in many countries, the U.S. version of which is whether the top individual income tax rate should be 35% as it is now, or 39.6% as it was in 2000 and will revert to in 2010 unless the law changes, or even higher.

However, two factors mitigate this policy conclusion. The first is that the tax-rate elasticity of current-year ordinary taxable income captures only a part of the relevant behavioral elasticities. In particular, it misses the substitution possibilities between ordinary income and capital gains and between current taxable income and future (i.e., deferred) taxable income. Thus, it is not sufficient to consider *the* elasticity of taxable income without considering at least two types of income and several periods, and the tax rates applied to the types of income and over time.

Among affluent taxpayers, opportunities to convert ordinary income into capital gains abound. A recently controversial example in the United States is the carried interest of private equity fund managers, whose typical 20% stake in the capital gains of the fund over a pre-specified hurdle rate receives capital gains tax treatment, although it is arguably the return to the effort and talent of the fund managers in identifying underperforming companies and turning them into more profitable enterprises. The incentive to convert ordinary income into capital gains depends, inter alia, on the tax rate differential between the two. Thus, increases in the ordinary income tax rate will increase the incentive to convert, as will decreases in the capital gains tax rate.

Nearly all of the ETI research has ignored the interaction between the ordinary rate and the capital gains rate, with most research addressing taxable income net of capital gains. For several reasons this methodological approach may provide misleading answers. First, to the extent that ordinary taxable income changes because the capital gains tax rate changes, the behavioral response may be misattributed to concurrent changes in ordinary tax rates or some other relevant factor. A similar misattribution will affect attempts to estimate the responsiveness of capital gains realizations to capital gains rates, if they ignore concurrent changes in ordinary income tax rates.

In all countries in which capital gains are preferentially taxed, which are almost all countries, the tax system must set limits on what qualifies as a capital gain. This means that, in the context of tax avoidance, a new concept is needed, which one might call a *tax line elasticity*, which summarizes the behavioral response to changing where the tax law draws the line between two differentially taxed activities, such as ordinary income and capital gains.

As I will expand on in a few minutes, lines are ubiquitous in tax law, and the crucial elasticities will in general depend on where these lines are drawn.

The avoidance opportunities available to the rich are heterogeneous. How easy it is to avoid or defer tax on income, or convert it into capital gains, depends on whether one is a CEO, an investment banker, an entrepreneur, or a professional basketball player. Not only do the avoidance opportunities differ among these groups, the non-tax objectives that have tax implications vary, too. For example, a CEO might seek to diversify her holdings of company stock or stock options, while minimizing the capital gains tax exposure and considering SEC insider trading and disclosure rules. An entrepreneur might seek to minimize the capital gains tax liability attendant to the sale of a business.

One key issue arises when private costs that constraint avoidance and evasion are not social costs. For example, if it is monetary penalties, i.e. fines, which constrain evasion, this is a private, but not a social cost, so that the ETI used for welfare analysis must be adjusted to reflect this. A recent paper by Raj Chetty clarifies that the key parameter for this adjustment is the relative contribution to the total marginal cost of reducing tax liability of, on the one hand, social costs and, on the other, what he calls transfer costs, i.e., costs that are private but not social costs. He suggests that the ratio of transfer costs to total marginal costs may be high, so that the social cost of taxing the rich may be considerably lower than the "rich ETI" might suggest. But the evidence for the quantitative importance of transfer costs is not yet strong, and measuring its role should be high on the empirical agenda of the analysis of tax systems.

The Endogeneity of Elasticities

The second important caveat to the naïve application of the ETI to policy arises because, when the relevant behavioral responses involve tax planning rather than labor-leisure choices, the elasticity is not immutable. Rather it is subject to policy manipulation by, for example, changing the tax base definition or changing the enforcement of existing law. We economists are accustomed to thinking of tastes, or utility functions, as being immutable, including but not restricted to people's tastes between leisure and market goods. However, once we admit other behavioral responses such as avoidance and evasion, we must address the fact that the behavioral elasticity is not necessarily immutable. In fact, it depends on a number of factors.

The tax system choices of the home government

The choices made by the government and its tax authority—the definition of the tax base, the remittance system and the enforcement system—affect the elasticity of response to tax rate changes. For example, the penalty for detected evasion affects the evasion component of the ETI, the rules regarding Cayman Islands bank accounts affect the ETI, as does how broad the income tax base is, because it determines the broadness of the set of untaxed alternatives.

Indeed, in principle, all tax system parameters affect the elasticity of response to a change in the tax rate. From this perspective, the key central ETI is not an exogenous parameter at all, but is the result of tax system choices that should themselves be optimized. Thus, we can think of there being an *optimal* ETI.

Moreover, the optimal setting of any one tax policy instrument depends on the setting of the others. The optimal graduation of the rate structure

spends on the setting of the other parameters of the tax system. Imagine if when Ronald Reagan was elected as U.S. president in 1980 he had commissioned a study of the ETI and, once having the results, concluded that the top tax rate was too high. It is possible, if the enforcement system was suboptimal, that this was the locally correct policy change but that the global optimum was to raise the top rate and beef up enforcement. Whether in 2009 the top U.S. federal income tax rate should go back up to its lofty Clintonian height of 39.6% from 35% depends on, for example, the rules regarding the tax treatment of the carried interest of private equity fund managers, because it is rules such as this that determine the "rich ETI".

Choices made by the tax business

A large industry exists to help taxpayers locate deductions and credits to which they are eligible, and to identify tax-saving activities that may be at the fuzzy border between legal and illegal, or even on the illegal side of any line. A sector of that industry, the tax shelter business, is in the business of innovating sets of transactions that, by combining tax code provisions and arbitraging inconsistent treatment of financial income, allow tax savings. The ideal tax shelter is one that reduces tax liability without requiring much, if any, distortion in the real activities of the company or individual and, for corporations, does not negatively impact financial statements—so that they reduce taxable income without lowering earnings. Much of the tax business is engaged in tax-driven product innovation, creating products, often financial products, that are just on the low-tax side of lines defined in the tax code on the basis of more or less observable characteristics. I will have more to say about lines later.

Tax Havens

The elasticity of taxable income may depend on the policies of other countries. A fascinating example of this is presented by tax havens, or in some instances of multilateral institutions. A tax haven is a jurisdiction that levies no or only nominal taxes and offers itself as a vehicle for non-residents to escape tax in their country of residence. A tax haven can offer this service because it has laws and administrative practices that prevent the effective exchange of information on taxpayers benefiting from the low-tax jurisdiction.

There is considerable concern that the havens are "parasitic" on the tax revenues of the non-haven countries, inducing them to expend real resources in defending their revenue base and in the process reducing the welfare of their residents. A 1998 OECD report concluded that "governments cannot stand back while their tax bases are eroded through the actions of countries which offer taxpayers ways to exploit tax havens [and preferential regimes] to reduce the tax that would otherwise be payable to them."

In sharp contrast to this longstanding concern about the deleterious effects of havens, recent normative economic theory has focused on a potentially beneficial role for tax havens. The starting point is the well-known result that, under certain conditions, a small open economy should levy no distorting tax on mobile factors such as capital. Countries do, however, levy distorting taxes on mobile capital, and much of the recent theoretical literature conceives of tax havens as a device to save these countries from themselves, by providing them with a way to move toward the non-distorting tax regime they should, but for some reason cannot, explicitly enact.

In a recent paper with Jay Wilson we develop a model of tax competition in the presence of tax havens that explains and justifies initiatives to limit haven activities. We model the decision of a country to become a haven and, in so doing, demonstrate that small countries have a greater incentive to become havens. The countries that choose to be havens are parasitic on the revenues of the latter, in the following sense. Tax havens are juridical entrepreneurs that sell to multinational corporations protection from home-country taxation, resulting in what some political scientists call the "commercialization of state sovereignty." They are, in essence, establishments in the "tax business." The equilibrium price for this service depends on the demand for such protection, which in turn depends on the tax system, including the resources devoted to tax enforcement by the non-haven countries, and on the technology available to the parasitic havens.

In the model, tax havens lead to the wasteful expenditure of resources, both by firms in their participation in havens and by governments in their attempts to enforce their tax codes. In addition, tax havens worsen tax competition problems by causing countries to further reduce their tax rates below levels that are efficient from the viewpoint of all countries combined. Either full or partial elimination of havens is found to be welfare-improving for the residents of non-haven countries. Most strikingly, initiatives to limit some, but not all, havens can be designed to make residents of *all* countries better off, including residents of the remaining havens, who can now receive more for their services due to restricted supply.

Investments in elasticity

A final reason for the endogeneity of the elasticity of taxable income is the behavior of the taxpayer. Whenever tax payments are the result of a

negotiation with the taxpayer, having options is favorable to the taxpayer. Multinational companies with the ability to shift real operations and taxable income abroad will be able to strike a better deal, and so will be more inclined than otherwise to invest to establish that flexibility. Even in the absence of bargaining power, with uncertainty about future policy taxpayers will want to have mobility options (akin to learning English in a foreign country). Furthermore, as tax rates rise, there are sectoral shifts toward difficult-to-tax things, like self-employment income, intangible capital, and mobile things. Because difficult-to-tax bases are generally more elastic, an increased tax rate endogenously increases the aggregate elasticity of taxable income.

Empirical Challenges with Estimating the ETI

The endogeneity of the ETI raises some difficult empirical challenges. In many cases when tax rates change, other aspects of the tax system such as the breadth of the tax system or the enforcement system also change. Thus we should expect that the ETI will differ before and after a reform, and so when the literature estimates *the* ETI, perhaps of a particular income group, is it estimating the pre-reform ETI, the post-reform ETI, or some linear combination that need not be bounded by either? More generally, one needs to carefully control for the policy and other factors that affect the ETI. Note that even without non-rate tax system policy changes, the argument I made above about sectoral shifts also renders the pre- and post-reform aggregate ETI different.

Firms and Remittance

As I have mentioned, almost all of modern tax theory is about what triggers tax liability. Who or what entity must remit the tax triggered is unspecified, and presumed irrelevant. As I have mentioned, there is an irrelevance proposition emphasized in all public finance textbooks that it doesn't matter which side of a taxed transaction must remit tax, the incidence is the same. So, for example, it doesn't matter if a retail business or a consumer remits the tax, the outcome is exactly the same. It doesn't matter whether only the retailer businesses remit, as under a RST, or whether all businesses remit on their value added, as in a VAT.

The theory of optimal commodity taxation reads as if consumers remit taxes, but they almost never do—firms do, either retail firms, as in a retail sales tax, or all firms, as in a value-added tax. As the RST or VAT suggest, individuals need not be involved at all in tax remittance/collection system. Even what is nominally a labor income tax need not involve individuals as remittors, as is the case with exact withholding systems or final withholding systems that are common in other countries.

The importance of firms to tax systems becomes apparent once one recognizes that it is cost-efficient for the tax authority to deal with a small number of entities with relatively sophisticated accounting and financial expertise rather than a much larger number of employees or providers of capital. The centrality of firms in remittance and information reporting is illustrated by two recent studies that find that, in both the United States and the United Kingdom, well over 80 percent of all taxes are remitted by business. Anecdotal evidence suggests that in developing countries the fraction of revenue collected from businesses is even higher.

Notably, though, dealing with *small* businesses is not generally cost-efficient, and many tax systems either entirely exempt small businesses from

remittance responsibility, or else feature special tax regimes for small businesses that simplify the tax compliance process, and thereby change the base on which tax liability is based. In many countries the exemption of small firms is *de facto*, due to ineffective enforcement, as in the US, where the IRS has estimated the small business non-compliance rate to be about two-thirds.

Although explicit or implicit exemption, or more generally special tax treatment, of small firms might economize on collection costs (both compliance costs borne in the first instance by taxpayers and administrative costs borne in the first instance by the tax authority), it also generally causes production inefficiency, in part because it provides a tax-related incentive for firms to be — or stay — small. The tradeoff between the costs of collection and production inefficiency has not been closely addressed by the optimal tax literature.

One reason for this lack of attention is that meaningfully heterogeneous firms are absent from the modern theory of taxation. Diamond and Mirrlees assumed constant-returns-to-scale technology for all firms in all sectors (or 100% tax on pure profits), which implied that firm size is indeterminate and irrelevant in the model. But it is not irrelevant in the world. In addition, recall that the famous Diamond and Mirrlees (1971) theorem on aggregate production efficiency demonstrates that production inefficiencies should not be tolerated if the government faces no constraints on its ability to levy optimal commodity taxes. But their model of optimal taxation ignores collection costs.

In recent work with Dhammika Dharmapala and Jay Wilson, we develop a model in which there are heterogeneous firm sizes generated by random draws of productivity parameters, and a fixed per-firm

administrative cost of having a firm in the tax net. The government must raise a fixed amount of net-of-cost revenue using three policy instruments: a constant tax rate on output, a fixed per-firm fee, and an output cutoff, below which firms are not taxed. While in the development literature entry fees have often been viewed as a manifestation of bureaucratic inefficiency or corruption, we show that when all firms in an industry are taxed, optimal policy may involve the use of the fixed fee; the fee basically acts like a Pigouvian tax, internalizing the social costs of tax administration. In our model, each *industry* is characterized by constant returns to scale, because the set of firms that are potential producers is effectively unlimited and *ex ante* identical. In this setting, the standard rules of optimal commodity taxation hold if there are no administrative costs, enabling us to isolate the implications of introducing these costs and the Diamond and Mirrlees theorem on aggregate production efficiency tells us that the tax system should not discriminate among firms in the same industry. With administrative costs, we identify conditions under which it is optimal to exempt small firms from taxation, thus creating production inefficiencies that are inconsistent with the optimal tax system in the Diamond and Mirrlees framework. These inefficiencies occur because different firms in the same industry sell output at different prices, and also because some firms obtain the tax exemption by reducing their outputs to inefficiently low levels, creating a "missing middle" of intermediate-sized firms that has been much discussed in developing countries. But this production inefficiency is balanced against the cost savings from collecting revenue from, on average, larger firms.

Before leaving the topic of the role of firms in tax systems, I want to note another great irony: that the discredited economic system of

communism had the most cost-efficient way of collecting taxes. This is not a bizarre coincidence. Government had control, indeed effective ownership, of all firms, the key to tax collection. Although the Soviet Union had an elaborate machinery of so-called taxes, this was a facade because the true tax burden—defined for labor income tax as the difference between the marginal product of labor and the employees' take-home pay—was almost entirely implicit. This was essentially a system of final, and invisible, employer withholding. The invisibility is part of the reason for the widespread antipathy to taxation in the post-Soviet era—employees had been unaware of their implicit tax burden, and were not used to dealing with a tax authority charged with collecting the true, previously implicit but now explicit, burden.

Line drawing

Real tax systems must also address line drawing. The real-world, in-the-trenches, scuffling about taxation, as any tax lawyer will know, is all about drawing and interpreting lines, yet analysis of this topic is completely absent from economic analysis. Why?

One reason is that the modern theory of optimal commodity taxation prescribes a different tax on each good, which depends on the nature of utility functions and perhaps also on distributional objectives and on the pattern of externality generation. But this is infeasible. Whenever selective commodity taxation is called for, a non-capricious tax system must have procedures for distinguishing among goods subject to different tax rates. Real-world consumption tax systems do that by appealing to the characteristics of the commodities. For example, the retail sales taxes of U.S. states often exempt food but not restaurant meals, requiring the tax law

to draw a line between the two categories. This is done by appealing to a set of characteristics of a restaurant meal, and the line can be fine when, for example, grocery stores sell pre-prepared meals that may or may not be eaten on the premises, or set up in-store salad bars. The retail sales tax in the Canadian province of Ontario exempts basic food items such as flour but applies to other processed foods such as chocolate bars, requiring lines to be drawn, including one that subjects to tax "biscuits or wafers specifically packaged and marketed to compete with chocolate bars." Several European countries provide a subsidy for certain kinds of consumer services (e.g., cleaning, gardening, and house repair) based on a Ramsey-type justification that such services compete with untaxed home production. This requires the classification of services eligible for the subsidy based on observable characteristics.

The prominent role of characteristics in commodity tax systems is due to several factors. First, the alternative that the theory implies—relying on estimates of the set of compensated elasticities—is infeasible. These elasticities are notoriously difficult to estimate precisely, and they would certainly not be intuitive to either policy makers or consumers in the way that characteristics-based rules are. Second, a shared characteristic plausibly signals something about the relative substitutability of the goods, and so may serve as a more readily measurable indicator of the ideal, but not observable, distinguishing factor. Third, modern economies produce a vast amount of different goods, and the set of available goods is constantly evolving. If tax laws were specified literally in terms of goods and their associated elasticities then, whenever a new good is introduced in the market, there would be no natural way to assign it to a tax category and the law would have to be re-specified to explicitly deal with the new good. In contrast, a

characteristics-based rule for assigning tax rates to goods naturally handles the creation of new goods by limiting the tax policy choice to which characteristic-based category the new good falls in.

In recent work with Henrik Kleven, we have been trying to re-formulate optimal commodity tax theory in the language of characteristics so that it matches up more easily with real tax systems. To do so we make use of Kelvin Lancaster's idea that is the characteristics of goods, not the goods themselves, which are the direct objects of utility, and there exists a mapping of each good into characteristics space. We formalize the relationship between characteristics, substitutability and optimal tax rates, which allows us to explore the notion that shared characteristics can be used to gauge substitutability and hence optimal tax rate differentials. We show that the closer two goods are in characteristics space, the smaller the optimal tax rate differential.

Once this reformulation is done, we can naturally address another important aspect of reality that has been ignored by the literature on optimal taxation: tax-driven product innovation. By this term we refer to the creation of new products that requires no technical innovation, but which represents a re-packaging of characteristics so as to reduce tax liability. For example, car manufacturers have an incentive to redesign vehicles to just qualify for gas-saving subsidies or just avoid gas-guzzler taxes. On Wall Street or the City of London, tax-driven product innovation is not a curiosum, but rather a major pre-occupation, where one objective is to design corporate finance vehicles that qualify for the interest deduction accorded to debt finance, but have most or all of the characteristics of an equity security.

In the standard optimal tax model, addressing the creation of new goods is not tractable, because a change in the set of available goods must be

associated with a new utility function (with new arguments) and therefore a new optimal tax problem. In the Lancaster approach, on the other hand, because the set of characteristics that consumers value is stable, the utility function is robust to the introduction of new goods and we can then incorporate product innovation into the optimal tax problem. We show that non-uniform tax systems may give rise to the creation of goods which are socially inferior in characteristics space, but which may be privately optimal for tax avoidance purposes. This represents a distortion in the set of available goods, which is different from the demand and supply distortions typically considered by public finance economists.

Furthermore, once we allow for the creation of new goods, it becomes clear that a tax system must include procedures for assigning potential (but currently non-existing) goods to tax categories. Much real-world tax legislation defines tax categories by listing a number of observable characteristics, and places any given commodity into the category with which it shares a majority of its characteristics, a procedure often called line drawing. Note that a "line" shares many attributes of a "notch" in tax schedules, which refers to a discontinuity in the function of how tax liability relates to the tax base, and which are generally not part of an optimal tax system. Indeed, *a line is a notch in characteristics space*, because the tax liability changes discontinuously as the characteristics vector of a good crosses the statutory line. Given our assumption that a continuum of tax rates is administratively infeasible, notches in characteristic space are an unavoidable feature of tax systems, not an idiosyncrasy.

We show in the paper that, under certain assumptions regarding the technology by which new goods can be created, the notches associated with line drawing create an incentive to the production and consumption of goods

that are just on the low-tax side of a line that separates two tax rate regions. We also demonstrate that, if administratively feasible, optimal lines are drawn so as to completely avoid tax-driven product innovation. In a world with just two goods and two tax rates, this implies that the line should be "close enough" to the characteristics of the low-tax good. This result may seem surprising at first glance in the sense that, even though we consider a second-best optimal tax problem, the solution ensures the existence in equilibrium of the first-best set of available goods. This is an unusual result when viewed from the perspective of the theory of second-best that prescribes that we typically do not want to completely eliminate the tax distortion on any given margin: we would rather have small distortions "everywhere" than large distortions somewhere and none elsewhere. The standard result is based on the notion that increasing a tax distortion around the point of no distortion is associated with only a second-order deadweight loss. But in our context, because of the unavoidable notch in goods creation, when the line is drawn so as to just allow for tax-driven product innovation to occur, a new good will be put on the market which will eliminate one of the existing goods. This creates a first-order welfare loss around the point of no distortion.

III. Conclusions

I've sketched out an agenda for making progress in the analysis of tax systems. If successful, it will build a bridge between rigorous analysis of taxation and the kind of tax system issues that are prominent in tax policy formulation. Although today I've stressed the theoretical issues, there is a parallel empirical agenda that focuses on refining measures of the responsiveness of the tax base to tax rate changes, not restricted to real

choice such as labor supply, and one that pays close attention to the effect of tax policy parameters other than rates and bases, including how they interact with rates to influence behavior, and to the sophisticated tax avoidance strategies that are available to the affluent and to multinational corporations.