Optimal Tax Remittance with Firm-Level Administrative Costs

Dhammika Dharmapala
University of Connecticut

Joel Slemrod
University of Michigan

John Douglas Wilson
Michigan State University

Working draft. Not for quotation.
Comments welcome.

October 30, 2007

1. Introduction

Firms play a central role in the remittance system of all modern taxes. This is largely because it is cost-efficient for the tax authority to deal with a small number of relatively financially sophisticated entities rather than a much larger number of employees or providers of capital. But dealing with *small* businesses is not generally cost-efficient, and many tax remittance systems either exempt small businesses from remittance responsibility, or else feature special tax regimes for the small business sector that simplify the 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 lax enforcement.[1] While this special treatment 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 an incentive for firms to stay small.

The tradeoff between collection cost savings and production inefficiency has not been closely addressed by the optimal tax literature. In part, this is because nearly all of modern tax theory is concerned with what actions or states of the world *trigger* tax liability, and virtually none is concerned with the system of remittance of funds to the government to cover that liability. Indeed, elementary public finance textbooks often assert that the remittance details — such as whether the buyer or seller of a commodity remits the sales tax triggered by the sale — are *irrelevant* to the consequences of a tax.[2] The operation of actual tax systems, however, requires considerable attention to the remittance of monies to the tax authority, including both the administration and enforcement of the tax rules, as well as the design of the tax rules with the administrative and compliance issues in mind.

The scant analytical attention paid to this issue is also in part due to the fact that that the canonical model of optimal taxation, represented by the seminal contribution of Diamond and Mirrlees (1971), makes assumption about technology (constant returns to scale or 100% profits tax) that renders firm size either meaningless or irrelevant. This

[1] For an excellent review of the sorts of special regimes that countries apply to small businesses, see IMF (2007).
[2] The relevance of the remittance system is discussed in Slemrod (2007).

assumption, while affording considerable analytical tractability and important insights,[3] is highly constraining in dealing with settings in which heterogeneous firm size, production inefficiency, and regressive collection costs of firm-based taxes, are ubiquitous features (see Kopczuk and Slemrod (2006)). This setting, we argue, is the one that arises in practice.

In this paper we contribute to the development of such a framework by addressing the properties of an optimal firm-based remittance system in the presence of exogenous per-firm administrative costs. We address the optimal properties both for a tax system whose instruments are a fixed per-firm fee[4] and a proportional output tax, and for a more general tax structure that allows for a fixed fee and a nonlinear tax on output. We develop a model in which each industry is characterized by constant returns to scale, because the firms available for production are effectively unlimited and ex ante identical. But after initially assuming that firms are also ex post identical, we develop models in which firms become differentiated in productivities after entering the industry. In this setting, the standard rules of optimal commodity taxation hold if there are no administrative costs, giving us an opportunity to isolate the pure effects of these costs. In particular, the Diamond and Mirrlees theorem on aggregate production efficiency tells us that the tax system should not discriminate among firms in the same industry. In contrast, we identify cases in which it is optimal to exempt small firms from taxation, thereby eliminating the administrative costs incurred in taxing them, even though the resulting differences in the prices at which small and large firms sell output creates the inefficiencies ruled out by the Diamond-Mirrlees theorem. Moreover, because the exemption is based on observed firm outputs, it causes some firms to avoid taxes by producing inefficiently low levels of output.

This latter distortion is emphasized by Keen and Mintz (2004), who also consider an output cutoff for exempting firms from a value-added tax in the presence of administrative and compliance costs. But their firms differ ex ante and earn untaxed

[3] See, in particular, the analysis of optimal taxation and administrative costs by Yitzhaki (1979) and Wilson (1989), which we later discuss

[4] IMF (2007, p. 31) discusses the fixed per-firm fee, also known as a *patente* system, as an example of a presumptive tax regime.

profits, in which case the Diamond-Mirrlees theorem no longer holds and different firms should generally be taxed at different rates, even without administrative costs.[5]

Heller and Shell (1971) provide an early analysis of the possibility that administrative costs may eliminate the desirability of aggregate production efficiency. But for most of the paper, they follow the prevailing practice of ignoring the relevance of the remittance system, and instead model the costs of commodity taxes as "indirectly accounted for through the consumer purchase and sales vectors" (p. 340). Towards the end of the paper, they do briefly associate administrative costs with individual firms (p. 344) and make the important observation that the government may want to shut down firms from which the collection of taxes is relatively costly, even if these firms are a little more efficient in production. In our model, if the government attempted to shut down firms producing low levels of output, then some of these firms would respond by inefficiently raising their outputs. But in our analysis, forced shutdowns of small firms are not needed, because the government can impose a fixed fee that introduces scale economies into their cost structures, causing the smaller firms to voluntarily shut down anyway.

We also develop rules for the relative levels of taxation in different industries. When there are untaxed firms, these rules include output supply elasticities, in contrast to the standard optimal tax rules. When the tax system consists of a per-firm fee and constant marginal tax on output, these elasticities include not only the price elasticity of supply for existing untaxed firms, but also an elasticity measuring the extent to which firms change from taxed to untaxed output as the price that firms receive for untaxed output rises. The presence of these elasticities tends to reduce the relative average tax on an industry's output, calculated net of administrative costs.

We begin in Section 2 with a model of identical firms, and show that the fixed fee should be set to equal the administrative costs, with any required revenue being raised with the proportional output tax. The fee basically acts like a "Pigouvian tax,"

[5] This discussion refers to the second of two models presented by Keen and Mintz. In the first model, firm sales are exogenous, so a cutoff rule does not distort output decisions. (See also Zee (2005) for an extension of this model.) Both models differ from our analysis by treating net product prices as fixed (a small open economy), assuming an exogenous social marginal value of government revenue, and not supplementing the VAT with fixed fees.

internalizing the social costs of tax administration.[6] The output tax should follow an inverse-elasticity rule, but with the added proviso that some goods should be completely exempt from taxation if the ratio of administrative costs to output tax payments exceeds a certain amount. Then in Section 3, we introduce heterogeneity in firm size that arises because of variations in productivity. This heterogeneity allows us to introduce the output cutoff, below which firms are not subject to tax but do not generate any administrative cost. In Section 4, we show that under certain conditions, cutoffs are part of the optimal tax structure. In Section 5, we introduce an optimal nonlinear tax on an industry's taxed output, emphasizing the use of the nonlinear structure to reduce taxes on firms that might otherwise avoid these taxes by inefficiently reducing their outputs to the cutoff level. In Section 6, we develop the modified inverse-elasticity rule for the average (net of administrative costs) taxes on different goods. Section 7 concludes, and an appendix provides the formal setup of the optimal nonlinear tax problem.

Taken together, the results of this paper suggest that in cases where it is not desirable to exempt whole industries from taxation, it may be desirable to take an intermediate approach by removing the tax burdens on relatively small firms within an industry and reducing the industry's overall level of taxation.

2. The Structure of Taxes across Industries with No Tax-Exempt Firms

In this section, differences across industries, rather than firms within an industry, are the major focus. Thus, let us first consider an economy with many industries, each of which has access to an unbounded mass of identical potential firms. Before making its entry decision, a firm knows that if it enters an industry, it will face a strictly increasing, strictly convex cost function given by $c_i(y_i) - c_{ei}$ for a firm with output y_i in industry i, where $c_i(0) = 0$ and c_{ei} is treated as an entry cost; thus, average cost curves are U-shaped. Free entry drives profits to zero, net of taxes. These taxes consist of a constant marginal tax rate on output, t_i, and a fixed fee, b_i. The fixed fee will turn out a critical

[6] If, instead of administrative costs, firms incurred fixed compliance costs in the payment of taxes, then the fee would not be needed because these costs would already be internal to the firm. Our other results are also easily modified to account for compliance costs.

component of the tax system when there are administrative costs in the collection of taxes. Letting p_i denote the producer price for good i, calculated net of the output tax, profit maximization yields a firm's output function, $y_i(p_i)$, and its profits function, defined gross of fixed costs:

$$\pi_i(p_i, b_i) = p_i y_i(p_i) - c_i(y_i(p_i)) - b_i . \quad (1)$$

Free entry reduces p_i to the point where the value of these profits just covers the fixed cost, c_{ei}.

The demand for each good is obtained from the utility-maximization problem for a representative consumer with utility function, $U(X, L)$, where X is a vector of I goods and L is the supply of an input called "labor." The representative consumer supplies this labor to firms in each industry and receives all profits, but we have noted that there are no profits in equilibrium. Labor serves as the numeraire, so the costs described above are measured in units of labor. To work with demand functions for each good that depend only on the good's own price, we assume that the direct utility function is quasi-linear in labor and separable. Letting q_i denote the price that the consumer pays for good i, utility maximization gives the demand functions, $X_i(q_i)$ for good i, and the indirect utility function, $\Sigma_i v_i(q_i)$. In the presence of a unit tax t_i on good i, this consumer price equals $p_i + t_i$. The number of firms producing good i, M_i, is determined by the requirement that demand equals supply: $X_i(q_i) = M_i \, y_i(p_i)$.

In this paper we focus on the implications for optimal tax policy of per-firm fixed costs of collecting taxes.[7] Specifically, the government incurs a fixed per-firm "administrative cost", A_i, when it collects taxes from a firm. This assumption is intended to capture the notion that there is a substantial fixed component to tax administration and enforcement. For instance, suppose that tax enforcement requires that (at least with some positive probability) the tax authority must dispatch agents to audit the records of each firm. Although it may be the case that auditing a larger firm requires more resources, it is exceedingly unlikely that such differences, even if they exist, will be proportional in size to the differences in firms' output levels. The assumption made here is that all

[7] We do not allow firms to split up or combine for tax purposes.

administrative costs are fixed and hence independent of the size of the firm; however, the qualitative results apply as long as there is a relatively large fixed component to the costs.

The government's tax instruments consist of the vectors of output taxes and fixed fees. The values of these control variables determine the market-clearing values of the consumer and producer prices, and the number of firms entering each industry. Following the standard practice in optimal tax theory, however, it is convenient to replace the tax rates with consumer and producer prices as control variables.[8] It is also convenient to make the M_i's control variables, subject to the constraint that they equate demand with supply in each product market. The government's objective is to maximize utility $\Sigma_i v_i(q_i)$. We next state the Lagrangian and then describe the constraints.

$$L = \sum_i v_i(q_i) + \sum_i \lambda_i \left(\mathrm{X_i(q)} - M_i \mathrm{y_i(p_i)}\right) \qquad (2)$$

$$+ \lambda_B \left(\sum_i \left(M_i \left((\mathrm{q_i} - \mathrm{p_i}) \mathrm{y_i(p_i)} + b_i - A_i \right) \right) - E \right) + \sum_i \beta_i \left(\pi_i(p_i, b_i) - c_{ei} \right).$$

The constraint multiplying the Lagrange multiplier λ_i is the requirement that demand equal supply in the market for good i. The multiplier λ_B multiplies the government budget constraint, where E is an exogenous revenue requirement. Finally, the multiplier β_i multiplies the requirement that profits equal zero in industry i.

We first show that the fixed fee is set to finance administrative costs, leaving the output tax to finance expenditure needs:

Proposition 1. *If all firms producing good i are taxed, then the optimal fixed fee equals administrative costs: $b_i = A_i$.*

Proof. Let $b_i > A_i$. Then lower $\mathrm{b_i}$ slightly, and offset the gain in profits, $M_i db_i$ by lowering p_i with q_i fixed (i.e., by levying a higher t_i): $- M_i db_i = X_i dt_i$. Then the zero-profit requirement remains satisfied. If there were no behavioral changes, we would then

[8] In constant-cost models, p_i is fixed by the technology, but we shall see that this is not the case here.

see that tax revenue stays the same. Demand X_i stays the same, since it is determined by q_i, which has not changed; but the fall in p_i lowers output per firm. As a result, more firms must enter the industry to keep total output equal to the fixed demand. None of these behavioral changes affect the zero-profit requirement, but revenue rises by $(b_i - A_i)dM_i > 0$. Thus, we have a revenue gain, and the surplus in the government budget can be used to lower q_i, raising welfare By reversing the argument, we find that b_i cannot be less than A_i. Thus, $b_i = A_i$. Q.E.D.

Proposition 1 can be understood as a marginal-cost-pricing requirement. A firm that chooses to produce output generates a "social cost" in the form of administrative costs, and the fee should internalize this cost.[9] This reasoning clearly does not depend on the simplifying assumption that all firms are identical, so Proposition 1 carries over to the model of heterogeneous firms described in the next section, assuming all firms are taxed.

With the fee taking care of administrative costs, we should expect the output tax to satisfy the usual inverse-elasticity rule for an optimal tax system.[10] This turns out to be the case. Let α denote the marginal utility of income, and define the price elasticity of demand as a positive quantity:

$$\varepsilon_i^X = -\frac{dX_i}{dq_i}\frac{q_i}{X_i}. \tag{3}$$

We then have:

Proposition 2. *If all firms producing good i are taxed, then the optimal output tax satisfies the inverse-elasticity rule:*

$$\frac{t_i}{q_i} = \frac{1-\frac{\alpha}{\lambda_B}}{\varepsilon_i^X}. \tag{4}$$

[9] See Slemrod (2007) for a discussion of the extent to which administrative and compliance costs are appropriately treated as externalities.
[10] See Auerbach and Hines (2002) for a careful exposition of this model.

Proof. Differentiating the Lagrangian with respect to M_i, we obtain a first-order condition that implies

$$\frac{\lambda_i}{\lambda_B} = \frac{R_i}{X_i} = t_i, \tag{5}$$

where R_i is the amount of revenue raised by taxes on firms in sector i, calculated net of administrative costs. Revenue per unit of output equals the output tax because administrative costs are financed by the fixed fee (Proposition 1). By differentiating the Lagrangian with respect to q_i and employing this equality, we obtain a first-order condition that implies (4). Q.E.D.

Although administrative costs do not affect the structure of output taxes on goods that are taxed, they do affect the set of taxed goods. Suppose, in particular, that for some reason administrative costs A_i incurred in taxing firms in sector i increase, while the costs related to another sector j (or set of other sectors) decrease, so that the government budget stays balanced with no change in taxes. Then (4) tells us that the optimal structure of output taxes does not change, and therefore social welfare stays constant, *if* we continue to tax the same goods. Instead, the government finances the higher A_i by raising the fixed fee b_i, and lowers b_j. If, however, A_i gets sufficiently high, then the government should reason that although its administrative costs are being paid for by the fixed fee, the fee is not generating any revenue net of administrative costs, whereas firms in industry i would greatly benefit from the elimination of their taxes, including the high fee. True, the revenue from the output tax, t_iX_i, would now have to be made up through higher taxes on the remaining taxed goods, and moving away from the tax structure dictated by the inverse-elasticity rule raises the deadweight loss from taxation. But the consumer price for good i, q_i, will have fallen so much to restore zero profits after the elimination of output taxes *and* the large fixed fee that consumers should still be better off.

In fact, if the initial output taxes are low relative to administrative costs, and a good is not such a large contributor to the government budget that exempting it would require much higher taxes on the remaining goods, then it should be exempted from taxation to obtain the savings in administrative costs. Using the quadratic formula for the deadweight loss of a tax, we next state a sufficient condition for this requirement that output taxes are low relative to administrative costs.[11]

Proposition 3. *Assume that all firms in all industries are initially taxed according to the inverse-elasticity rule, and no industry is large enough to generate more than a small share of total tax revenue. Then social welfare can be increased by exempting good j from taxation if the ratio of total administrative costs to total output tax payments for good j satisfies*

$$\frac{M_j A_j}{t_j X_j(q_j)} > \frac{\frac{t_i}{q_i}\varepsilon_i^X}{1-\frac{t_i}{q_i}\varepsilon_i^X}\left(\frac{1}{1-\frac{t_i}{q_i}\varepsilon_i^X}-.5\right), \tag{6}$$

where i is any taxed good other than j and linear approximations are used to measure the effects of taxes on demands.

Proof. Consider eliminating the output tax on some good j, causing a net loss of tax revenue equal to $t_jX_j(q_j)$. Then the consumer price drops from q_j by the amount t_j to the producer price p_j (see Figure 1). The gain in consumers' surplus exceeds the lost tax revenue:

$$\Delta_t CS_j \equiv \int_{p_j}^{q_j} X_j(p_j)dp_j = t_j X_j(p_j)\left(1+.5\frac{t_j}{q_j}\varepsilon_j^X\right), \tag{7}$$

[11] Condition (6) is a sufficient condition, but it is not a necessary condition because it ignores some of the gains in consumers' and producers' surplus to arise from the elimination of the fixed fee [see (10)].

which is equal to tax revenue plus the deadweight loss from the tax, using the quadratic expression for the deadweight loss. Eliminating the fixed fee results in a further price drop to a lower consumer price, q_j*, to keep profits equal to zero.

$$M_j \int_{q_j*}^{p_j} y_j(p_j)dp_j = M_j A_j . \tag{8}$$

Equation (8) gives

$$(p_j - q_j*)y_j(p_j{}^{\#}) = A_j \tag{9}$$

for some $p_j^{\#}$ between p_j and q_j*. The increase in consumers' surplus resulting from this price reduction satisfies

$$\begin{aligned} \Delta_a CS > (p_j - q_j*)X_j(p_j) &= \frac{M_j A_j}{M_j y_j(p_j{}^{\#})} X_j(p_j) \\ &= \frac{M_j A_j}{M_j y_j(p_j{}^{\#})}\left(X_j(q_j) - t_j X_j'(q_i)\right) > M_j A_j\left(1 - \frac{t_j}{q_j}\varepsilon_j^X\right). \end{aligned} \tag{10}$$

The first inequality holds because the change in consumers' surplus includes the expansion in output beyond $X_j(p_j)$ (see Figure 1), the first equality uses (9), the second equality uses a linear approximation, and the last inequality holds because $M_j y_j(p_j{}^{\#}) < M_j y_j(p_j) = X_j(q_j)$.

On the other hand, only the lost revenue from the output tax needs to be recouped through a rise in the taxes on the remaining goods; the decline in fixed fees is exactly offset by a decline in administrative costs. Assuming this revenue loss is small relative to total tax revenue, it will not matter which taxes on other goods are increased to offset the revenue loss, as long as these tax increases are kept small rather than being focused on a small group of other goods (in which case first-order approximations don't work because the tax changes are large). Thus, the loss in consumers' surplus from raising another dollar of revenue is:

$$\frac{dCS_i}{dR_i} = \frac{X_i}{X_i + t_i(dX_i / dq_i)} = \frac{1}{1 - \frac{t_i}{q_i}\varepsilon_i^X}, \tag{11}$$

where i is any taxed good other than j. Multiplying by the lost revenue, $t_jX_j(q_j)$, and comparing to the increase in consumers' surplus given by the sum of (7) and (10) then gives the following condition for an increase in welfare:

$$1 + .5\frac{t_j}{q_j}\varepsilon_j^X + \frac{M_j A_j\left(1 - \frac{t_j}{q_i}\varepsilon_j^X\right)}{t_j X_j(q_j^*)} > \frac{1}{1 - \frac{t_i}{q_i}\varepsilon_i^X}, \tag{12}$$

which can be rewritten as (6). Q.E.D.

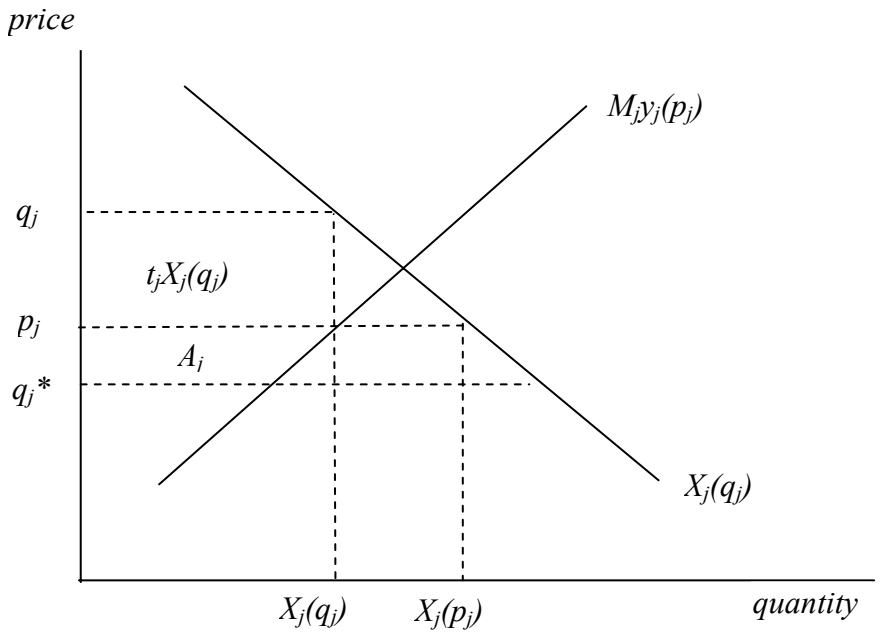

Figure 1

To illustrate the implications of (6), it is useful to consider a numerical example. Note that the term, $\frac{dCS_i}{dR_i} = \frac{1}{1 - \frac{t_i}{q_i}\varepsilon_i^X}$, is the marginal cost of public funds, a concept that has been widely studied. Although there is a wide range of empirical estimates, 1.25 is a reasonable midpoint of the serious estimates. This value implies that $\frac{t_i}{q_i}\varepsilon_i^X = .2$, which would hold under an elasticity of one and a twenty percent tax rate. If we interpret this tax rate as the one that would prevail when all goods are taxed, then it should be viewed as lying below the tax rates under the tax structures that countries typically use, because they tend to exempt many commodities from taxation. With this value for the marginal cost of public funds, condition (6) becomes

$$\frac{M_j A_j}{t_j X_j} > .1875 .$$

This implies that it will be desirable to exempt a good from taxation if its administrative costs are more than about twenty percent of the value of its output tax revenue (not including the fixed fees). It is interesting that this figure is less than the amount by which the deadweight loss from taxation is assumed to raise the marginal cost of public funds. Thus, these calculations suggest that if administrative costs are at least as important as deadweight losses considerations, then a good should be exempted from taxation. For industries consisting of relatively small firms, they are likely to be more important.

Proposition 3 has interesting parallels to the work of Yitzhaki (1979) and Wilson (1989). These papers consider the optimal taxation of a continuum of goods that enter a representative consumer's utility function symmetrically. Taxing each good involves a fixed administrative cost that varies across goods, but the standard deadweight loss from taxation rises as more goods (the ones with relatively high administrative costs) are exempted from taxation and taxes are increased on the remaining goods to balance the government budget. A similar tradeoff is at work in this paper, and we could analyze the optimal number of taxed goods in an economy with symmetric sectors. As the number

of untaxed goods rose, the increased taxes on the remaining taxed goods would raise the marginal efficiency losses from taxing them.

In the Yitzhaki and Wilson models, the reason that administrative costs vary across goods is exogenous. Nor could the source be related to administrative costs at the firm level, as it is assumed that all firms exhibit constant returns to scale, so the size of each firm is indeterminate. In this setting, with a fixed cost of tax administration for each good, if the government levied a fixed fee, then a competitive equilibrium would not exist, because the fee would introduce increasing returns to scale into firm activities, providing incentives for firms to grow without bound (until they recognized their influence over prices and stop behaving competitively).

In our model, U-shaped average cost curves limit the equilibrium size of firms. Heterogeneity of production technologies introduce heterogeneity across sectors in the costs of collecting taxes, as it is more costly to collect taxes from industries whose technology favors small firms, even when the tax authority optimally uses the policy instruments at its disposal (including both the fixed fee that mirrors the fixed per-firm administrative costs, and, as modeled below, a threshold size for being subject to tax).

In this manner, we endogenize interindustry differences in administrative costs and obtain a tradeoff between these costs and the deadweights costs of taxation that is similar to the tradeoff studied by Yitzhaki (1979) and Wilson (1989).[12] As evidenced by Proposition 3, industries with many small firms (high M_i's) are likely to exhibit relatively high administrative costs in tax collection, increasing the likelihood that they should be exempted from taxation. But raising the tax rate on a good will eventually cause (6) to be violated. It will not be optimal to exempt any firms from taxation if the tax payments collected from each industry are sufficiently high relative to administrative costs.

3. A Model with Heterogeneous Firms

Within an industry, firms typically differ significantly in size. For example, small grocery stores and large supermarkets both sell food. Given our assumption that

[12] The per-firm fixed fee may create incentives for firms to merge, depending on the extent to which the merged firm faces decreasing returns to scale in production. We leave an analysis of such incentives to future research.

administrative costs have an important fixed component, it follows that the government might wish to exempt small firms in an industry from taxation, but not large firms. We next develop a model that can be used to investigate this possibility. The model is inspired by Hopenhayn (1992a, b) and Melitz (2003)**,** but has been simplified to de-emphasize the dynamics that are a central focus of those papers. It makes use of the competitive setting used by Hopenhayn (1992a, b), rather than the monopolistic competition model of Melitz (2003).

After describing the behavior of firms, we will describe the government's optimal tax problem. Since we are interested in how firms within a given industry should be taxed, we drop subscripts identifying goods and focus on a single industry. In particular, we assume that the government has chosen the consumer price for this industry, q, fixing demand at $X(q)$, and we solve the suboptimization problem of maximizing net revenue, given q. If revenue were not maximized, then it would be possible to move to a different tax system that created a budget surplus that could be passed on to consumers through a welfare-improving reduction in q. The solution to this optimal tax problem will involve excess burden considerations and also administrative costs. Subsequently, we derive a modified inverse-elasticity rule for how the consumer prices on different goods should be chosen.

Building on our previous model, assume again that all firms in the industry face the same fixed cost, c_e, but assume that the convex variable cost function differs across firms: $c(y,\varphi)$ for a type-φ firm, where φ is an unknown productivity parameter that takes on values over an interval, $[\varphi^l, \varphi^h]$.[13] The value of φ cannot be discovered unless the firm enters, although its distribution (characterized by cdf $F(\varphi)$ and pdf $f(\varphi)$) is known *ex ante,* and there is an unbounded mass of identical potential entrants, each possessing this distribution. Each firm chooses its output only after incurring the fixed cost, c_e, which is unrecoverable. We assume that a higher value of φ decreases marginal costs: $\frac{\partial c(y,\varphi')}{\partial \varphi} < \frac{\partial c(y,\varphi'')}{\partial \varphi}$ for $\varphi' > \varphi''$ and any value of y, including $y = 0$. Thus, a firm's chosen output rises with φ, and *ex post* profits also increase with φ. For a firm facing

[13] Unless otherwise stated, φ^l can be taken to be minus infinity, and φ^h plus infinity.

unit output price p and fixed fee b, these profits, calculated gross of the fixed entry cost, are given by

$$\pi(p,b,\varphi) = py(p,\varphi) - c(y(p,\varphi),\varphi) - b\,, \tag{13}$$

where the supply function, $y(p, \varphi)$, is obtained by maximizing profits.

Once a firm has entered the industry, it incurs no costs if it does not produce (i.e., $c(0, \varphi) = 0$) and may therefore avoid any additional costs by exiting immediately. Thus, it will exit if the φ it draws is too small to yield non-negative profits; that is, if $\pi(p, b, \varphi) < 0$. Exit entails no costs, but the entry cost c_e is sunk and cannot be recovered. For the subsequent analysis, we assume that there always exist sufficiently unproductive firms for exit to occur.

Let us now introduce a threshold y^* for taxed outputs. Firms with outputs equal or less than y^* are not taxed and thus receive the consumer price q per unit of output. Firms with outputs above y^* pay the fixed fee b and are taxed at the rate t on their output, in which case they face the producer price $p = q - t$ per unit of output. The assumption here is that the government is unable to observe productivities directly, and must therefore base its tax on observed outputs. If the output cutoff is high enough to induce a positive measure of firms to produce untaxed output, then it easy to see that the optimal tax system will cause some of these firms to be bunched at y^*, and the minimum output of taxed firms will lie above y^* by some discrete amount. If there were no bunching, the government would be incurring administrative costs in collecting negligible amounts of taxes from firms producing outputs slightly above y^*, in which case it would be desirable to raise y^* or change the tax system to induce bunching.

For a given cutoff level, y^*, the degree to which firms bunch at y^* can be reduced by raising the output tax and lower the fixed fee, without necessitating a change in q, since those taxed firms that produce relatively low levels of taxed output (i.e., values of $y(p, \varphi)$ near y^*) are affected relatively little by the higher output tax but benefit from the lower fee. On the other hand, the higher output tax distorts output decisions of those firms that pay the tax. Given these conflicting considerations, we are unable to sign the difference between b and A when there is an output cutoff, but this sign is not needed for our subsequent results.

Let $\pi^*(q, y^*, \varphi)$ denote the profits received by a bunched firm, which face the consumer price q:

$$\pi^*(q, y^*, \varphi) = qy^* - c(y^*, \varphi) \ . \tag{14}$$

At the productivity level where a firm is indifferent between paying the tax and reducing output to y^* to avoid the tax, we then have[14]

$$\pi^*(q, y^*, \varphi) = \pi(p, b, \varphi). \tag{15}$$

This equality defines a unique cutoff productivity, $\varphi^{**}(q, p, b, y^*)$. Firms with a productivity φ' above φ^{**} choose not to reduce output to y^*, because the drop in taxes at y^* (from $ty^* + b$ to zero) is offset by a greater loss in profits from the requirement that output be reduced to y^*. In contrast, all firms with productivities below $\varphi^{**}(q, p, b, y^*)$ but above some $\varphi^*(q, y^*)$ choose to produce output of y^*. Combining these observations gives:

$$y = y^* \quad \text{for} \quad \varphi^*(q, y^*) < \varphi < \varphi^{**}(q, p, b, y^*); \tag{16}$$

$$y = y(p, \varphi) \quad \text{for} \quad \varphi > \varphi^{**}(q, p, b, y^*). \tag{17}$$

If there are firms producing positive outputs below y^*, then $\varphi^*(q, y^*)$ is found by inverting $y(q, \varphi) = y^*$. But for a sufficiently low value of y^*, it may the case some firms choose y^*, but no firm can make positive profits by operating below y^*, in which case $\varphi^*(q, y^*)$ solves $\pi^*(q, y^*, \varphi) = 0$. Otherwise, the minimum productivity at which profits are non-negative satisfies $\pi(q, 0, \varphi) = 0$. To deal with both contingencies, define $\varphi^m(q, y^*)$ as this minimum productivity for an economy with price q and cutoff y^*, recognizing that $\varphi^m(q, y^*) \leq \varphi^*(q, y^*)$, with strict inequality if y^* is sufficiently high.[15] Then the total output supplied by untaxed firms consists of a mass of outputs that each equal y^*, and a distribution of outputs below y^*, if any exist:

[14] Note that π^* and π are different functions, with different arguments.
[15] In the absence of a cutoff, the minimum productivity would be a function of the producer price p and fixed fee b. We are assuming here that the cutoff is high enough to exempt some firms.

$$Y^U = M\left(\int_{\varphi^m(q,y^*)}^{\varphi^*(q,y^*)} y(q,\varphi)f(\varphi)d\varphi + \int_{\varphi^*(q,y^*)}^{\varphi^{**}(q,p,b,y^*)} y^* f(\varphi)d\varphi\right), \tag{18}$$

where M remains the number of firms entering the industry. The output of taxed firms is

$$Y^T = M\left(\int_{\varphi^{**}(q,p,b,y^*)}^{\phi^h} y(p,\varphi)f(\varphi)d\varphi\right). \tag{19}$$

M is determined by the requirement that total supply equal the fixed demand:

$$Y^U + Y^T = X(q). \tag{20}$$

In addition, the expected profits for firms entering the industry equals zero.

This and other equilibrium requirements are included in the Lagrangian for the government's problem of maximizing revenue from this industry, given the price q:

$$L = M\left[\int_{\varphi^{**}(q,p,b,y^*)}^{\varphi^h} \big((q-p)y(p,\varphi)+b-A\big)f(\varphi)d\varphi\right] \tag{21}$$

$$\lambda\left(X(q) - M\left[\int_{\varphi^m(q,y^*)}^{\varphi^*(q,y^*)} y(q,\varphi)f(\varphi)d\varphi + \int_{\varphi^*(q,y^*)}^{\varphi^{**}(q,p,b,y^*)} y^* f(\varphi)d\varphi + \int_{\varphi^{**}(q,p,b,y^*)}^{\phi^h} y(p,\varphi)f(\varphi)d\varphi\right]\right)$$

$$+\ \beta\left(\int_{\varphi^m(q,y^*)}^{\varphi^*(q,y^*)} \pi(q,0,\varphi)f(\varphi)d\varphi + \int_{\varphi^*(q,y^*)}^{\varphi^{**}(q,p,b,y^*)} \pi^*(q,y^*\varphi)f(\varphi)d\varphi + \int_{\varphi^{**}(q,p,b,y^*)}^{\phi^h} \pi(p,b,\varphi)f(\varphi)d\varphi - c_e\right).$$

The Lagrange multiplier, λ, multiplies the market-clearing constraint, and the Lagrange multiplier, β, multiplies the zero-profit constraint. These constraints account for the three types of firms: those below the cutoff (if there are any), which pay no taxes; those at y^*, which also pay no taxes; and those that produce in excess of the cutoff and are therefore taxed on their output. The control variables are the producer price p, which determines the tax $t = q - p$; the fee, b; the number of entrants, M; and the cutoff, y^*. The first-

order condition for M shows that the Lagrange multiplier is the ratio of revenue, net of administrative costs, to output, which we denote T^e:

$$\lambda = \frac{R}{X} \equiv T^e . \tag{22}$$

4. Should there be a Cutoff?

If the tax authority can exempt from tax firms below a certain size, should it? Some initial intuition may suggest that the answer is yes. If a firm's output is small, little revenue, net of administrative cost, is obtained from taxing it. In particular, the fixed fee is covering administrative costs (Proposition 1), and the tax on the low level of output yields little revenue. Thus, taxes on the other taxed firms will not need to be increased much to offset the revenue loss from exempting the small firm from taxation. But whereas the government cares about net tax payments, the firm benefits from the elimination of its gross tax payments, much of which consists of the fee. Without these gross tax payments, the firm's profits rise by a sizable amount, so either the consumer price will fall to restore zero expected profits for firms entering the industry, or the government can raise taxes on other firms in the industry without causing the consumer price to rise. In other words, society benefits from the exempting the firm from taxation.

An immediate problem with this intuition is that the requirement that firms effectively finance the costs of tax administration by remitting the fixed fee means that they cannot be too small if administrative costs are sizable, because they must earn sufficient amounts of revenue to finance the fee. Thus, the fee basically increases the minimum size of firms in the industry, and so the optimal minimum size of firms remitting the tax may encompass all firms.

Another problem is that the output cutoff will induce some firms to escape the tax net by producing less than their most efficient output levels, resulting in a deadweight loss that may offset the benefits of an output threshold. In fact, this consideration becomes so important when the government attempts to exempt only a small number of firms from taxation that welfare necessarily declines. We now state and prove this negative result.

Proposition 4. *Starting from a welfare-maximizing tax system without output cutoffs, introducing a cutoff for firms in a given industry must lower welfare if the resulting set of untaxed firms is sufficiently small.*

Proof. If we set y^* at the highest point where there is no bunching ($\varphi^{**} = \varphi^*$), only type-φ^{**} firms are willing to produce untaxed output (and they are indifferent). Differentiating the Lagrangian given by (21) for the revenue-maximization problem then gives

$$\frac{\partial L}{\partial y^*} = T^e(y^{**} - y^*)\frac{\partial F^{**}}{\partial y^*} - \left(\mathrm{t}y^{**} + b - A\right)\frac{\partial F^{**}}{\partial y^*}, \tag{23}$$

where use is made of (22) and $\partial F^{**}/\partial y^*$ is the marginal rise in the share of firms that produce untaxed output. Evaluate this derivative at $b = A$, as required by Proposition 1 for the initial equilibrium without untaxed output:

$$\frac{\partial L}{\partial y^*} = t(y^{**} - y^*)\frac{\partial F^{**}}{\partial y^*} - ty^{**}\frac{\partial F^{**}}{\partial y^*} = -ty^*\frac{\partial F^{**}}{\partial y^*} < 0\,. \tag{24}$$

Thus, starting from the highest cutoff at which there are no untaxed firms, raising the cutoff creates a first-order revenue loss for a fixed value of q, implying that welfare must fall. Q.E.D.

The basic idea here is that if only a few firms can gain by choosing to produce untaxed output, then they cannot benefit much. Otherwise, there would be many other firms also benefiting from producing untaxed output. The benefits of the tax exemption for these few firms must be offset by the profit losses they incur to reduce their outputs to untaxed levels. (The derivative, $\frac{\partial L}{\partial y^*}$, reflects this absence of a profit gain.) Thus, the movement of firms to untaxed output is not generating a rise in expected profits for the industry, and the government is therefore unable to raise taxes on firms above the cutoff

without necessitating a rise in the consumer price to satisfy the zero-expected-profit requirement. But with some output now produced by untaxed firms and total output fixed at *X(q),* there must be a reduction in total *taxed* output, even after we account for the entry of firms into the industry needed to keep supply equated with demand after some existing firms reduce their outputs to y^*. This decline in the tax base lowers tax revenue, necessitating tax increases to balance the government budget, and the consumer is harmed by a higher consumer price.[16]

But as the cutoff level is increased, more firms take advantage of it, and firms that are initially at the cutoff level see their profits rise. Thus, the cutoff does start to raise expected profits, enabling the government to increase its taxes on firms above the cutoff without causing the consumer price to rise. Thus, we are able to prove:

Proposition 5. *For a given level of administrative costs, if the tax revenue collected from an industry in excess of these costs is sufficiently small in the absence of an output cutoff, then the welfare-maximizing tax system will involve a cutoff that is high enough to induce a positive measure of firms to produce untaxed output.*

Proof. Returning to the Lagrangian given by (21) for the revenue-maximization problem, set $b = A$ (as required by Prop. 1 in the absence of a cutoff) and $t = 0$, and differentiate the Lagrangian with respect to y^*:

$$\frac{\partial L}{\partial y^*} = \beta \int_{\varphi^*}^{\varphi^{**}} \left(q - c_y(y^*, \varphi)\right) f(\varphi) d\varphi \ , \tag{25}$$

where use is made of (22). As y^* increases above the highest level where no firm benefits from producing y^* ($\varphi^{**} = \varphi^*$), firms bunch at y^* ($\varphi^{**} > \varphi^*$) and further increases in y^* raise profits for these bunched firms, because they sell at a price q above their marginal costs at y^*. Thus, the expected profits available to a firm entering the

[16] We could eliminate this revenue loss by changing p and b in a way that eliminated incentives for firms to switch to untaxed output y^*. But this would violate the condition in the proposition that some firms do produce untaxed output.

industry rise. Equation (25) measures the marginal value of the profits generated by a unit rise in y^*. By raising y^* to generate these profits, the government can then satisfy the zero-expected-profit requirement by raising the tax t and fee b. Thus, government revenue rises, implying a welfare gain. Since revenue rises when $t = 0$ at the no-cutoff optimum, the continuity properties of the model imply that revenue will also rise when this t is positive but sufficiently small. Q.E.D.

The basic argument here may be simply explained. Increasing the cutoff generates additional profits for existing untaxed firms, allowing the government to raise taxes on firms above the cutoff without necessitating a rise in the consumer price to keep expected profits equal to zero. But a higher cutoff reduces taxable output, which tends to reduce tax revenue. If the average net tax on output, T^e, is small enough, however, then this second effect will be unimportant, and so additional revenue can be generated by raising the cutoff, at no cost to the consumer. This additional revenue can then be distributed to the consumer through a reduction in the consumer price.

Note that this welfare gain requires that the cutoff be high enough to induce a sizable number of firms to produce untaxed output. Otherwise, Proposition 4 applies. But as T^e goes to zero, the initial welfare losses from raising the cutoff above the point where it begins to attract firms go to zero and can be overcome by subsequent welfare gains.

5. Nonlinear Tax Systems

As evidenced by Proposition 4, much of the potential benefit of an output cutoff for taxation may be offset by its negative impact on firms' output decisions, as they attempt to qualify for tax-exempt status by reducing their outputs to inefficiently low levels. One way to counteract these inefficiencies would be to give tax breaks to firms with taxable outputs near the cutoff. Simply stated, if tax burdens were lowered on firms that might be tempted to lower their outputs to the cutoff level, then they might decide not to do so. This variation in outputs could be achieved using a nonlinear tax system based on output, under which a net tax function, $T(y)$, is chosen so that firms choosing a

relatively low output pay a relatively low average net tax, $T(y)/y$, where net again means after administrative costs are subtracted. In this section, we provide a simple rule for taxing a firm that is indifferent between taxed and untaxed output. We then provide a rule for the optimal cutoff under a nonlinear tax. Finally, we discuss possible shapes of the entire nonlinear tax schedule over taxed output levels. In particular, we argue that to keep the average tax rate relatively low at output levels near the cutoff, the marginal tax should be set so that firms with higher outputs pay higher average taxes.

We again focus on a single industry and consider the suboptimization problem of choosing the cutoff and nonlinear tax schedule to maximize tax revenue, given the chosen consumer price, q, which determines demand, $X(q)$, and welfare. The cutoff at y^* already gives us a special type of nonlinearity in the tax system, with $T(y) + A = 0$ for $y \leq y^*$; that is, there are no gross tax collections. But now this nonlinearity is supplemented by marginal taxes, dT/dy, that vary with income at $y > y^*$.

We leave the formal treatment of the optimal tax problem to an appendix, and instead conduct our analysis with the aid of Figure 2, which presents a possible tax schedule, beginning at y*. Profit indifference curves over tax payments and output are drawn for type-φ^* and type-φ^{**} firms, where these two types are the endpoints of the interval of firms that bunch at y^*. Any given firm would want to choose an output on the lowest possible of its profit indifference curves, subject to being on the tax function. The type-φ^* firm maximizes profits at y^*, with or without taxes, and the type-φ^{**} firm maximizes profits at both y^* and y^{**}. We have constructed the tax schedule to rise above the type-φ^{**} firm's indifference curve as output increases from y^{**}, so that the firm will not choose any output between y^* and y^{**}. But we could clearly eliminate the bunching of firms at y^* by replacing the kink in the tax schedule at y^* with a smooth tax schedule. But then there would exist some firms producing slightly above y^* and remitting almost no taxes, while the government incurred the administrative cost A to collect these taxes. Hence, the government would want to change the tax function so that those firms that do produce above y^* make tax payments that are large enough to justify the required administrative costs.

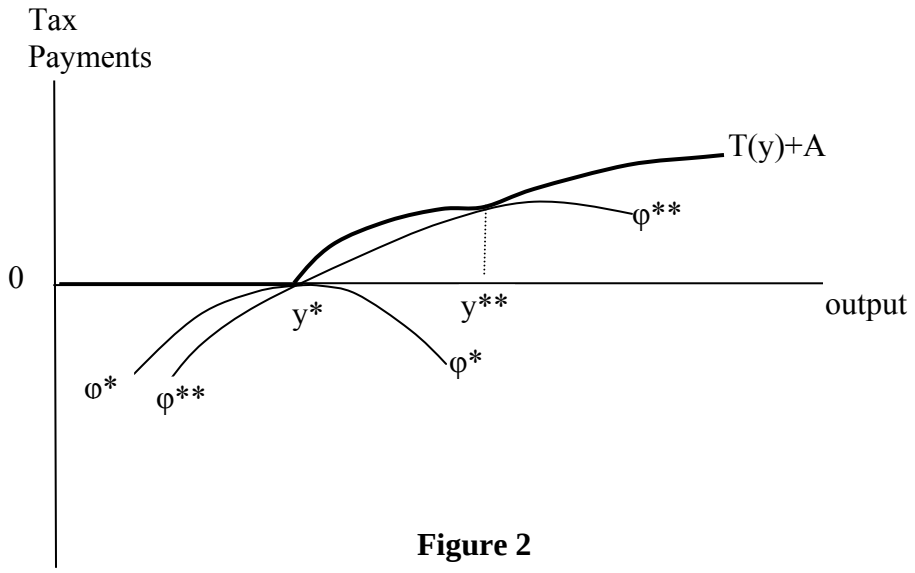

Figure 2

Consider now the rule for the optimal number of firms to exempt from taxation. The critical insight here is that this number can be increased by raising tax payments by a small amount in a small interval of incomes from y^{**} to $y^{**} + dy^{**}$. Since firms in this interval are approximately indifferent between y^* and y^{**}, this change causes a small number of firms to switch from taxed output to untaxed output, without significantly altering their profits. The direct revenue loss per firm is $T(y^{**})$, calculated net of the savings in administrative costs. Offsetting this loss, however, is the gain in revenue from an increase in entry into the industry. Since total demand stays fixed at $X(q)$, an existing firm's decision to reduce output from y^{**} to y^* must be offset by a change in the number of firms entering the industry, dM, that satisfies, $y^e dM = y^{**} - y^*$, where y^e is the expected output per firm entering the industry. Thus, $dM = \frac{y^{**} - y^*}{y^e}$. Following our previous notation, again let T^e be the tax per unit of output for each firm entering the industry, calculated net of administrative costs. Then the tax payments per firm are $T^e y^e$. Multiplying this amount by the change in M gives the total change in tax payments from this entry: $T^e(y^{**} - y^*)$. For the initial tax system to maximize revenue, this revenue gain must exactly offset the net loss in revenue, $T(y^{**})$, that directly resulted from the firm's

switch from taxed output to untaxed output. Thus, we have the following rule for the optimal number of firms to exempt from taxation:

$$T(y^{**}) = T^{e}(y^{**} - y^{*}). \tag{26}$$

Thus, the tax system is optimal only when there is no net change in tax revenue generated by a marginal firm's switch from taxed to untaxed output. An interesting aspect of this rule is that it contains no terms involving the behavioral responses of firms to tax changes. It is the rule that the government would want to follow to maximize tax payments if it had chosen an output cutoff y^* and could directly control the number of firms that produce at y^* or y^{**}. This direct control is effectively available to the government through its choice of the marginal firm's tax payments. In contrast, the government can alter the number of taxed firms in the linear tax model only by changing the common tax rate t, which causes all taxed firms to alter their chosen output supplies.

Using (26), we next obtain an intriguing condition that must be satisfied if the cutoff y^* is also optimized. Note that $q - c_y(y^*, \varphi)$ may be viewed as a type-φ firm's implicit marginal tax on output, since it is the marginal tax that would cause the firm to reduce output to y^* in the absence of a cutoff rule. With this interpretation, we have:

Proposition 6. *If an optimal nonlinear tax system includes an output cutoff, then it causes enough firms to bunch at the cutoff, y^*, to equate the average of their implicit marginal taxes on output to the industry's average net tax on output:*

$$\frac{\int_{\varphi^*}^{\varphi^{**}} \left(q - c_y(y^*, \varphi)\right) f(\varphi) d\varphi}{\int_{\varphi^*}^{\varphi^{**}} f(\varphi) d\varphi} = T^{e}. \tag{27}$$

Proof. We first derive an optimality condition for y^* in the linear-tax case that is also applicable to the nonlinear-tax case. In particular, consider marginal increase in y^* that

is accompanied by a rise in the fee b that keeps expected profits equal to zero. Let dF^{**}/dy^* denote the resulting change in the share of firms that are untaxed. This derivative is positive because the higher y^* makes untaxed output more profitable, and the higher b makes taxed output less profitable. For the initial y^* and b to be optimal, the derivative of the Lagrangian in (21) with respect to y^* and this change in b must equal zero:

$$\frac{\partial L}{\partial y^*} + \frac{\partial L}{\partial b}\frac{db}{dy^*} = T^e\left[(y^{**} - y^*)\frac{dF^{**}}{dy^*} - \int_{\varphi^*}^{\varphi^{**}} f(\varphi)d\varphi\right] \tag{28}$$
$$+ \left[-T(y^{**})\frac{dF^{**}}{dy^*} + \int_{\varphi^*}^{\varphi^{**}}\left(q - c_y(y^*,\varphi)\right)f(\varphi)d\varphi\right] = 0,$$

where $T(y^{**}) = ty^{**} + b - A$. But the same policy perturbation that gives us (28) is also available in the nonlinear case: raising b by a unit is equivalent to increasing $T(y)$ a unit at each taxed y. In both cases, the marginal taxes on output (t and dT/dy in the linear and non-linear cases, respectively) do not change.

Using (26), (28) reduces to (27). Q.E.D.

This average implicit marginal tax measures the marginal impact of an increase in y^* on the average profits of taxed firms. These higher profits enable the government to raise its taxes on taxed firms without violating the zero-expected-profit condition. But this must be balanced against a shrinking tax base, which causes a revenue loss that depends on T^e on the right side. Eventually, the profit gain must fall below T^e because, as y^* increases, marginal cost must rise to price q for each bunched firm, eliminating the implicit tax. In other words, the left side of (27) must eventually fall below the right side, implying a limit on how much the cutoff output should be increased. Note, however, that (27) does not allow us to conclude that a sufficiently low cutoff output improves welfare. As discussed in the section on linear taxation, the problem is that any potential welfare gains from cutoffs that attract only a small number of firms are wiped out by the distorting impact of the cutoff on the outputs of these firms. In the nonlinear

tax case, the government responds to a welfare-reducing cutoff by collecting relatively low taxes from the marginal firms, in which case these firms are not willing to lower their output much below y^{**} to eliminate taxes. With very little difference between y^{*} and y^{**}, the left side of (27) is close to zero and therefore less than T^e, consistent with the negative impact of increasing y^{*} on welfare.

In the appendix, we derive tax schedules for firms above the cutoff y^{**} that possess the form shown in Figure 3, which depicts both the marginal and average net tax rates as functions of output. In particular, the average net tax for a firm with the lowest taxed output lies below the average net tax (i.e., $T(y^{**})/y^{**} < T^e$), a property that follows directly from (26). But the marginal tax on this output is high and declining, eventually falling below the average tax as the top output is approached.[17]

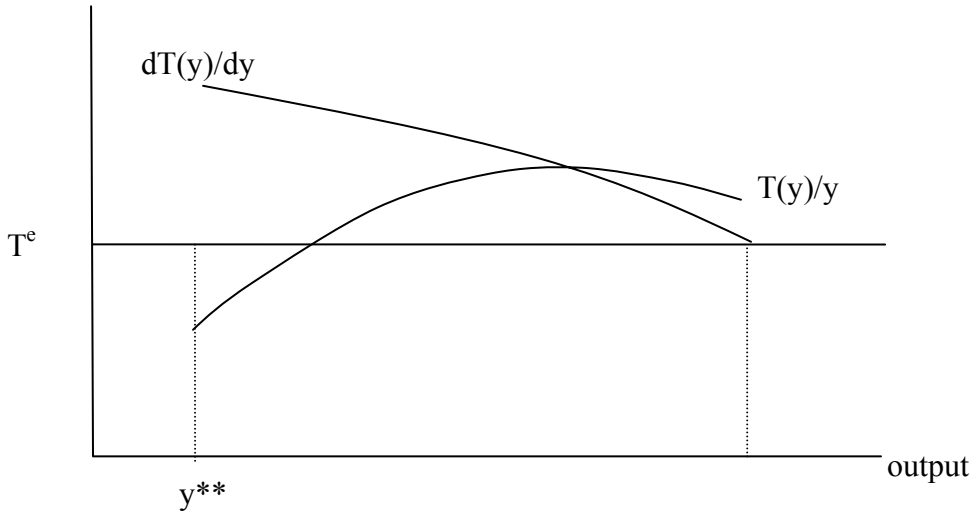

Figure 3

While allowing nonlinearity in this tax system clearly can mitigate the distortions created, this would just as clearly involve additional administrative costs in practice. For this reason, the restriction to linear taxes should be taken seriously. Nevertheless, the analysis in this section suggests the desirability of limited nonlinearities in the tax system, whereby the government couples the elimination of taxes at sufficiently low

[17] See Proposition A in the Appendix. We do not show that $T(y)/y$ always monotonically declines with y, but it must lie above T^e and decline to T^e at the top output level under the conditions stated in the appendix.

outputs tax with breaks for somewhat higher output, but limits the number of marginal tax brackets at higher outputs.

6. A Modified Inverse-Elasticity Rule

When there are untaxed firms, the previously-derived inverse-elasticity rule no longer holds. In that rule, only the demand elasticity mattered, because raising the consumer price alone had no impact on the producer price received by firms, the level of which is fixed by the requirement that expected profits equal zero. But untaxed firms receive the consumer price, so supply elasticities matter. To see this, define good *i*'s supply elasticity of untaxed output as follows:

$$\varepsilon_i^U = \frac{\partial Y_i^U}{\partial q_i}\frac{q_i}{Y_i^U} > 0\,. \tag{29}$$

For this definition, we hold fixed the number of firms producing untaxed output and consider only the marginal impact of the price they receive (i.e., the consumer price) on their supplies. A change in q_i and any accompanying changes in the other tax variables will generally change the share of firms that produce untaxed output, denoted F_i**. Considering first the case of linear taxation, define $\frac{dF_i\,**}{dq_i}$ as the change in this share as a result of a marginal rise in q_i, accompanied by a "profit-preserving" rise in the fixed fee, b_i. This derivative is positive, because the higher consumer price makes untaxed output more attractive, and the higher fee makes taxed output less attractive. Changing this share will alter tax revenue by

$$\frac{dR_i}{dF_i\,**} = M_i\left(T_i^e\,(y_i\,** - y_i\,*) - T_i\,**\right). \tag{30}$$

T_i** is the loss in revenue from the movement of a firm from taxed output y_i** to untaxed output y_i*, and $T_i^e\,(y_i\,** - y_i\,*)$ is the resulting rise in revenue from the resulting entry of additional firms to increase total output back to total demand. For the case of nonlinear taxation, we proved in the previous section that this derivative is equal to zero.

But in the case of linear taxation, we shall show below that revenue falls. Define the following revenue and share elasticities:

$$\varepsilon_i^R = \frac{dR_i}{dF_i^{**}} \frac{F_i^{**}}{R_i} < 0; \qquad \varepsilon_i^F = \frac{dF_i^{**}}{dq_i} \frac{q_i}{F_i^{**}} > 0. \tag{31}$$

Again let α denote the consumer's marginal utility of income, and λ_B denote the marginal value of government revenue. We are now ready to derive a modified inverse-elasticity rule:

Proposition 7. *Under an optimal linear tax system, the average net tax rate for each taxed good satisfies the following modified inverse-elasticity rule:*[18]

$$\frac{T_i^e}{q_i} = \frac{1 - \frac{\alpha}{\lambda_B}}{\varepsilon_i^X + \frac{Y_i^U}{X_i}\varepsilon_i^U - \varepsilon_i^R \varepsilon_i^F} \quad . \tag{32}$$

Proof. To derive the new rule, recall the previous suboptimization problem: given a good's consumer price, the government maximizes the revenue obtained from taxing the good. Let $R_i(q_i)$ denote this maximized value of revenue for good *i*.. This function will depend on whether linear or nonlinear tax schedules are used, and on the use of a cutoff rule. The optimal tax problem then consists of maximizing the indirect utility function, subject to the government budget constraint:

$$Max \ \ \Sigma_i \, v_i(q_i)$$

$$s.t. \ \Sigma_i \, R_i(q_i) = E,$$

where E is, as before, the government's revenue requirement. Using Roy's Identity, the first-order condition for q_i is

[18] We limit this rule to "taxed goods" because we have seen that not all goods are necessarily taxed in the presence of administrative costs.

$$-\frac{\alpha}{\lambda_B} X_i(q_i) = \frac{dR_i(q_i)}{dq_i}. \tag{33}$$

To obtain a modified inverse-elasticity rule, we therefore need to calculate the revenue derivative. By the envelope theorem, this derivative is simply the derivative of the Lagrangian for the revenue-maximization problem. Also according to the envelope theorem, this derivative will not depend on whether we also change the fixed fee or the producer price as the consumer price rises, because both have a zero marginal impact on the Lagrangian at the optimum.

Omitting subscripts to simplify notation, let us then raise q while also increasing the fixed fee by an amount, db/dq, that keeps expected profits equal to zero: $db/dq = Y^U/(M(1-F^{**}))$. Differentiating the Lagrangian given by (21) with respect to q and this change in b, and using the previously-derived equality, $\lambda = T^e$, gives

$$\frac{dR}{dq} = X + T^e \frac{dX}{dq} - T^{**}M \frac{dF^{**}}{dq} - T^e\left(\frac{\partial Y^U}{\partial q} - (y^{**} - y^*)M \frac{dF^{**}}{dq}\right), \tag{34}$$

where T^{**} denotes the net tax payments for the type-φ^{**} firm (i.e., $T^{**} = ty^{**} + b - A$) and $\frac{dF^{**}}{dq}$ is the derivative of the share of firms that are untaxed with respect to q and the "profit-preserving" rise in b.

Using the expression given by (30) for the revenue derivative, $\frac{dR}{dF^{**}}$, we may rewrite (34) as

$$\frac{dR}{dq} = X + T^e\left(\frac{dX}{dq} - \frac{\partial Y^U}{\partial q}\right) + \frac{dR}{dF^{**}}\frac{dF^{**}}{dq}. \tag{35}$$

Substituting this derivative into (33) and expressing the result in elasticity form yields (32). Q.E.D.

This is a rule for the average net tax on output as a percentage of the consumer price—an average ad valorem net tax. Because not all output is taxed and the fixed fee

need no longer equal per-firm administrative costs, this average tax need no longer equal the constant marginal tax on output, *t*. In fact, we can demonstrate that, subject to an innocuous assumption,[19] that the marginal tax exceeds the average tax, in which case we can also show that the moving a firm from taxed output to untaxed output lowers revenue:

Proposition 8. *Under an optimal linear tax system with given cutoffs for each taxed industry, if* $(y_i^{**}-y_i^{*}) \le X_i/M_i$ *in taxed industry i, then the optimal taxation of that industry satisfies,* $t_i > T_i^e$ *and* $\varepsilon_i^R < 0$.

Proof. Again omitting subscripts, hold *q* fixed and raise *p* (i.e., a lower *t*) while increasing the fixed fee *b* to again keep expected profits fixed. For the initial *p* and *b* to be optimal, the resulting change in the Lagrangian, given by (21), must equal zero at the optimum:

$$\frac{\partial L}{\partial p}+\frac{\partial L}{\partial b}\frac{db}{dp}=T^eM\left[(y^{**}-y^{*})\frac{dF^{**}}{dp}-\int_{\varphi^{**}}^{\varphi^h}\frac{\partial y(p,\varphi)}{\partial p}f(\varphi)d\varphi\right] + M\left[t\int_{\varphi^{**}}^{\varphi^h}\frac{\partial y(p,\varphi)}{\partial p}f(\varphi)d\varphi-(ty^{**}+b-A)\frac{dF^{**}}{dp}\right]=0. \tag{36}$$

The change in the share of firms that are untaxed, dF^{**}/dp, is positive because although expected profits have not changed, the rise in *p* and *b* have increased the tax burden of type-φ^{**} firms, since their relatively low output y^{**} implies that they benefit relatively little from the higher *p*.

Suppose that

$$t \le T^e \equiv t\frac{Y^T}{X}+\frac{(b-A)M}{X}. \tag{37}$$

[19] The condition, $(y^{**}-y^{*}) \le X/M$, states that the gap between the cutoff output, y^{*}, and the lowest taxed output, y^{**}, is no greater than the average output over all firms in the industry. This is a sufficient, but not necessary, condition for (39) to hold. It seems improbable that the jump from y^{*} to y^{**} will large enough to violate this condition when the cutoff is optimized, but we do not have a proof ruling it out.

Then (36) gives

$$T^e M\left[(y^{**}-y^*)\frac{dF^{**}}{dp}\right] - M\left[(\mathrm{ty}^{**}+b-A)\frac{dF^{**}}{dp}\right]\geq 0\,. \tag{38}$$

Substituting the expression for T^e given by (37) into (38) yields

$$t\left(\left(\frac{Y^T}{X}\right)(y^{**}-y^*)-y^{**}\right)+(b-A)\left(\frac{M(y^{**}-y^*)}{X}-1\right)\geq 0. \tag{39}$$

The first of the two terms in (39) is clearly negative. Since (37) also implies that $b > A$, the assumption that $(y^{**} - y^*) < X/M$ implies that the second term is non-positive. Thus, (39) is violated, allowing us to conclude that $t > T^e$. Then (36) implies the remaining part of the proposition. Q.E.D.

It follows from this proposition that the modified inverse-elasticity rule for average net tax rates places a lower bound on the marginal tax rates on output, whereas the marginal and average rates (t and T^e) are equal in the no-cutoff case. To reduce the wasteful bunching of firms at the cutoff, y^*, the government can raise t above T^e and use the revenue to lower the fixed fee. While the lower fee benefits all taxed firms equally, the higher output tax burdens firms at the margin (type-φ^{**}) less than other taxed firms, since the marginal firms' outputs are relatively low. This argument does not necessarily imply that the fee should be reduced below administrative costs; indeed, doing so might require an excessive taxation of the reduced value of taxable output. But setting t above T^e does limit the amount by which the fee can exceed administrative costs.

The modified inverse-elasticity rule tells us that not only should taxes be low on goods with high demand elasticities (as the Ramsey rule instructs), but other things equal, they should also be low on goods with high supply elasticities for untaxed firms. The basic idea is that taxing output at a higher rate distorts not only demand decisions, but

also distorts supply decisions by increasing untaxed output at the expense of taxed output. The supply elasticity reflects this latter distortion, because the positive impact of a higher consumer price on the supplies of existing untaxed firms crowds out taxed output, through a reduction in the number of firms entering the industry. But the higher consumer price reduces taxed output through a second avenue: firms producing taxed output find untaxed output more attractive, causing some of them to switch. The share elasticity accounts for this latter consideration, and its importance depends on the revenue elasticity, which we have found to be negative: the movement of another firm from taxed to untaxed output lowers revenue.

For any given good, an increase in the magnitude of any of the elasticities just described will tend to lower the average net tax, relative to other goods. In addition, we see that goods with relatively high shares of output that are untaxed will have relatively low average taxes, all else equal. The reason is that the supply elasticity for untaxed output becomes more important in the rule as the untaxed-output share rises.

Suppose that the government uses a cutoff for some industries but not others. In fact, since Proposition 7 does not actually require that each industry's cutoff be optimally set, we could consider the case where no cutoff is used for industries consisting of a relatively small number of large firms, whereas industries containing a large number of small firms are subject to a uniform cutoff. The following proposition then follows:

Proposition 9. *If taxed goods i and j have the same demand elasticity, but the government uses a cutoff rule only for i under an optimal linear tax system, then the optimal average net tax on output is lower for i than for j:* $\frac{T_i^e}{q_i} < \frac{T_j^e}{q_j}$.

Proof. For those industries without the cutoff, only the demand elasticity enters the rule in Proposition 7. With the cutoff, the other elasticities enter and all contribute to a lower tax. Q.E.D.

An intriguing aspect of this result is that it is independent of how the administrative cost parameter varies across industries. But these costs do enter the net average tax rate as a sum (see (37)). Presumably the cutoff rule is used in industries with relatively high administrative costs. If this is the case, Proposition 9 tells us that the industries with the high administrative costs tend to have the lower average net taxes. In other words, additional gross tax payments are not fully covering the higher administrative costs.

Total administrative costs will tend to be relatively large for industries whose technologies dictate that they will consist of relatively small firms. To the extent that the government responds by using a cutoff rule for industries populated by small firms, but not those with large firms, Proposition 9 also suggests that low taxes should be levied on small firms, net of these administrative costs. Of course, their gross taxes might be high to cover these administrative costs.

Turn now to the use of nonlinear taxes. We argued previously that the rule for y^* under a linear tax system also applies to a nonlinear tax system, because the government can adjust the tax system as y^* rises in the same way: raise taxes so that each taxed firm's profit falls by the same amount (equal to the rise in the fixed fee in the linear case). Similarly, Propositions 7 and 9 also apply to the nonlinear case, except that (26) and (30) show us that the revenue elasticity is now zero, thereby eliminating the term in Proposition 7 involving the share elasticity. Thus we have:

Proposition 10. *Under an optimal nonlinear tax system, the average net tax rate for each taxed good satisfies the following modified inverse-elasticity rule:*

$$\frac{T_i^e}{q_i} = \frac{1 - \dfrac{\alpha}{\lambda_B}}{\varepsilon_i^X + \dfrac{Y_i^U}{X_i}\varepsilon_i^U}$$

We emphasize that this rule does not require that the entire nonlinear tax schedule be optimal. Rather, we are using only the optimality rules for firms at the margin between taxed and untaxed output. In addition, the government could respond to the

higher administrative costs associated with nonlinear tax structures by using nonlinear taxes for some, but not all, goods. In this case, Propositions 7 and 10 tell us that, all else equal, goods facing nonlinear taxes should be taxed more heavily in terms of average net tax rates than goods facing linear taxes. This is not surprising, since the use of nonlinear taxes must reduce the distorting effects of raising revenue.

As a point of comparison, one might consider a tax exemption threshold based on productivity rather than output. We have not up to now addressed this case because of the presumption that a firm's productivity is not observable by the tax authority. Nevertheless, it is of interest to note that this case gives exactly the same rule that appears in Proposition 10, because the potential distortion that it eliminates is the ability of firms to eliminate their taxes by reducing their outputs. In the nonlinear tax case, firms have this opportunity, but (26) implies that the welfare effect is zero at the margin.

An alternative possibility would be to randomly select firms to exempt from taxation. Proposition 10 would hold here too, again because output decisions are no longer being distorted by a cutoff rule. An interesting question is when the output cutoff should be chosen over random selection as a means of exempting firms from taxation. Presumably the output cutoff is superior in cases where firms differ significantly in size, with many small firms existing at the lower end of the distribution.

7. Conclusions

To be relevant to a world (like ours) in which there are significant tax administrative and compliance costs, a theory of optimal tax systems must address who or what entities remit tax as well as what triggers a tax. This requires attention to the role of firms in tax systems, and the difficult problem of collecting tax from small firms.

We develop models that produce some insight into the optimal design of tax systems when there are significant fixed per-firm costs of collection. These models constitute a bridge to Yitzhaki (1979) and Wilson (1989), who assume that there is an exogenous distribution of collection costs across goods, and that a sector may be either exempt from taxation completely or taxed uniformly. We provide an explanation for varying collection costs based on differing technologies that generate different size

distributions of firms across industries. These differences, in turn, render more or less important the fixed per-firm collection costs. In this setting, we generalize the policy instruments available to the tax authority by allowing them to collect taxes from some, but not all, firms in a sector. In addition, we show that a fixed per-firm fee is an essential component of an optimal tax system, and we show that average taxes on output should differ across industries, depending not only on the elasticity of demand for the good (as in a standard Ramsey model), but also on the size distribution of firms and the supply responses of firms to a tax increase (which raises the prices received by untaxed firms).

The models we have developed in this paper also show that optimal remittance systems generally induce production inefficiency. This is in contrast to the well-known finding of Diamond and Mirrlees (1971) that under certain assumptions including the absence of administrative costs, an optimal tax system will always satisfy aggregate production efficiency.

Future modeling work might usefully focus on some aspects of this model that are highly stylized. For example, the models in his paper presume that all goods are produced in a single stage of production. Thus it cannot address the important production efficiency questions that rise in the analysis of cascading business turnover taxes (also known as gross receipts taxes). Nor can it address without some refinement the issue that arises in a value-added tax that exemption of firms that sell to non-exempt firms (or to exempt firms that sell to non-exempt firms, etc.) does not provide a preference, and induces the formation of "chains" of exempt firms selling to each other and ultimately to final consumers.

Appendix

In this appendix, we first set up the nonlinear tax problem, then we obtain the first-order conditions, and finally we state and prove a proposition showing that the relation between a firm's tax payments and its output has the properties illustrated in Figure 3 in the text.

We again consider the problem of maximizing a single industry's tax revenue, given the choice of consumer price q. After solving this problem for each industry, the optimal q's can be determined in the manner described in Section 6.

As the control variable for the optimal tax problem, we can use the output cutoff y^*, the productivity that divides taxed and untaxed firms, previously denoted φ^{**}, the outputs for each taxed firm, $\{y(\varphi)\}$, the net profits for each taxed firm, $\{\pi(\varphi)\}$, and the number of firms that enter the industry, M. The government controls these net profits by taxing these firms. If $T(\varphi)$ is the firm's tax payments, calculated net of administrative cost A, then $T(\varphi)+A$ is its actual payment of taxes, and

$$\pi(\varphi) = [qy(\varphi) - c(y(\varphi), \varphi)] - (T(\varphi)+A). \tag{A.1}$$

We do not have to treat $T(\varphi)$ as a separate control variable, because (A.1) determines it as a function of the other control variables: $T(\varphi) \equiv T(\pi(\varphi), y(\varphi), \varphi)$. We omit q as an argument of functions because it is fixed for this analysis.

The productivity φ^{**} equates profits $\pi(\varphi^{**})$ under taxed output $y(\varphi^{**})$ to profits if the firm produces untaxed output y^*:

$$\pi(\varphi^{**}) = \pi^*(y^*, \varphi^{**}). \tag{A.2}$$

Since no taxes are paid at y^*, $\pi^*(y^*, \varphi) = qy^* - c(y^*, \varphi)$.

The government's objective is to maximize the sum of net tax payments across all firms. As before, the constraints include the requirement that the supply of output equals the fixed demand, $X(q)$, and the requirement that expected profits equals zero. But now there is an additional incentive-compatibility constraint, which we next discuss.

The government offers firms with outputs above y^* a menu of output-tax combinations, $\{y(\varphi), T(\varphi)\}$, which is designed so that a type-φ firm prefers output $y(\varphi)$

and tax $T(\varphi)$ over the other combinations of outputs and taxes. The first-order condition that must hold for this choice to be optimal for the type-φ firm is:

$$[q - c_y(y(\varphi), \varphi)]y'(\varphi) - T'(\varphi) = 0, \tag{A.3}$$

where a prime denotes a derivative. It follows from this first-order condition that

$$\pi'(\varphi) = - c_\varphi(y(\varphi), \varphi), \tag{A.4}$$

that is, the total derivative of profits with respect to the productivity parameter equals the partial derivative.

To incorporate these incentive-compatibility constraints into the Lagrangian, introduce the Lagrange multiplier $\mu(\varphi)$ for the constraint applying to type-φ firms. The Lagrangian will then include the following expression for these constraints:

$$\int_{\varphi^{**}}^{\varphi^h} \mu(\varphi)\left(\pi'(\varphi) + c_\varphi(y(\varphi),\varphi)\right)d\varphi . \tag{A.5}$$

where φ^h is again the top productivity. Integration by parts gives

$$\mu(\varphi^h)\pi(\varphi^h) - \mu(\varphi{**})\pi(\varphi{**}) - \int_{\varphi^{**}}^{\varphi^h}\left(\mu'(\varphi)\pi(\varphi) - \mu(\varphi)c_\varphi(y(\varphi),\varphi)\right)d\varphi \tag{A.6}$$

or, using (A.2),

$$\mu(\varphi^h)\pi(\varphi^h) - \mu(\varphi{**})\pi^*(q, y^*, \varphi{**}) - \int_{\varphi^{**}}^{\varphi^h}\left(\mu'(\varphi)\pi(\varphi) - \mu(\varphi)c_\varphi(y(\varphi),\varphi)\right)d\varphi . \tag{A.7}$$

With this form of the incentive-compatibility constraints, the Lagrangian may be written

$$L = M\left[\int_{\varphi^{**}}^{\infty} T(\pi(\varphi), \mathrm{y}(\varphi), \varphi) f(\varphi) d\varphi\right] \tag{A.8}$$

$$+ \lambda\left(X(q) - M\left[\int_{\varphi^m}^{\varphi^*} y^u(\varphi) f(\varphi) d\varphi + \int_{\varphi^*}^{\varphi^{**}} y^* f(\varphi) d\varphi + \int_{\varphi^{**}}^{\varphi^h} y(\varphi) f(\varphi) d\varphi\right]\right)$$

$$+ \beta\left(\left[\int_{\varphi^m}^{\varphi^*} \pi^u(\varphi) f(\varphi) d\varphi + \int_{\varphi^*}^{\varphi^{**}} \pi^*(y^*, \varphi) f(\varphi) d\varphi + \int_{\varphi^{**}}^{\varphi^h} \pi(\varphi) f(\varphi) d\varphi\right] - c_e\right)$$

$$+ \mu(\varphi^h)\pi(\varphi^h) - \mu(\varphi^{**})\pi^*(q, y^*, \varphi^{**})$$

$$- \int_{\varphi^{**}}^{\varphi^h} \left(\mu'(\varphi)\pi(\varphi) - \mu(\varphi) c_\varphi(y(\varphi), \varphi)\right) d\varphi$$

where φ^m is again the lowest productivity among firms that choose to produce, φ^* is the productivity that separates firms producing untaxed output y^* from firms producing less than y^* (if there are any), and the functions $y^u(\varphi)$ and $\pi^u(\varphi)$ are the output and profit functions for firms not subject to taxation and not bunched at y^* (in which case they are not control variables for the government). Having analyzed the choice of y^* in the text, we will keep it fixed for the analysis and instead focus on tax systems that are optimal, given the choice of y^*. Thus, φ^m can be treated as fixed. Marginal changes in φ^* do not affect the Lagrangian and so do not show up in the subsequent first-order conditions.

Turning to the first-order conditions, note first that $\pi(\varphi^h)$ now serves as a control variable, and differentiation gives

$$\mu(\varphi^h) = 0; \tag{A.9}$$

that is, the incentive-compatibility constraint does not bind at the top productivity level. Next, use the first-order condition for M to obtain:

$$\lambda = R/X \equiv T^e \tag{A.10}$$

where R is total revenue. This condition is also derived in the text for the linear-tax case. Using it, we can differentiate the Lagrangian and simplify to obtain the following first-order condition for φ^{**}:

$$T^e\left[y(\varphi^{**}) - y^*\right] - T(\varphi^{**}) = 0\,, \tag{A.11}$$

which is derived in the text using a less formal argument. Note that (A.11) implies that

$$\frac{T(\varphi^{**})}{y(\varphi^{**})} = T^e\,\frac{y(\varphi^{**}) - y^*}{y(\varphi^{**})} < T^e\,. \tag{A.12}$$

In words, the average net tax on output for a firm at the margin between taxed and untaxed output is less than the average net tax for the industry as a whole. To make up for this discrepancy, the average net tax must eventually rise above T^e at sufficiently high values of y, which means an increasing marginal tax. Figure 3 is consistent with these properties.

The first-order condition for $\pi(\varphi)$ is

$$-\,Mf(\varphi) \;+\; \beta f(\varphi) \;-\mu'(\varphi) = 0, \tag{A.13}$$

where the first term reflects the tax function derivative, $T_\pi(\pi, y, \varphi) = -1$. By integrating the left side and using our previous finding that $\mu(\varphi^h) = 0$, we obtain a formula for $\mu(\varphi)$:

$$\mu(\varphi) = \int_{\varphi}^{\varphi^h}\left(Mf(\varphi) - \beta f(\varphi)\right)d\varphi\,, \tag{A.14}$$

which may be written in terms of the distribution function, $F(\varphi)$, as follows:

$$\mu(\varphi) = \left(M - \beta\right)\left(1 - F(\varphi)\right). \tag{A.15}$$

Using the equalities, $\lambda = T^e$ *and* $T_y(\pi, y, \varphi) = q - c_y$, the first-order condition for $y(\varphi)$ may be written:

$$-T^e + (q - c_y(y(\varphi),\varphi)) + \frac{\mu(\varphi)}{Mf(\varphi)} c_{\varphi y}(y(\varphi),\varphi) = 0 . \qquad \text{(A.16)}$$

Combining the last two equations then gives

$$(q - c_y(y(\varphi),\varphi)) - T^e = \frac{(1 - F(\varphi))}{f(\varphi)} c_{\varphi y}(y(\varphi),\varphi)\left(\frac{\beta}{M} - 1\right) = 0 . \qquad \text{(A.17)}$$

Since firms marginal-cost price, $q - c_y(y(\varphi),\varphi)$ equals the marginal tax on output, denoted *dT(y)/dy* in Figure 3 in the text.

Using these first-order conditions, we now prove a proposition giving the properties of the tax schedule that are illustrated in Figure 3. To state the proposition, let $y^h = y(\varphi^h)$ and above y** = y(φ**). For parts of the proposition, we also assume:

Assumption A. $\lim_{\varphi \to \varphi^h} \frac{1 - F(\varphi)}{f(\varphi)} = 0 .$

In words, $(1\text{-}F(\varphi)/f(\varphi)$ goes to zero as φ goes to φ^h. This condition clearly holds for a uniform distribution. The proposition can now be stated as follows:

Proposition A. *Let the function T(y) describe the optimal relation between a firm's tax payments and its output. Then:*

(a) the marginal tax, dT(y)/dy, lies above T^e for $y^{**} \le y < y^h$;

(b) under assumption A, , the marginal tax falls to T^e at y^h;

*(c) the average tax, T(y)/y, lies below T^e at y = y**, then rises above T^e at higher output levels;*

*(d) under assumption A, this average tax reaches a maximum at some y between y** and y^h.*

Proof. By assumption A, (A.17) gives

$$(q - c_y(y(\varphi^h), \varphi^h)) - T^e = 0, \tag{A.18}$$

which is part (b).

Note that $c_{y\varphi}$ in (A.17) is always negative by assumption: higher-productivity firms have lower marginal costs. In then follows from (A.17) that at all outputs below $y^h = y(\varphi^h)$ and above $y^{**} = y(\varphi^{**})$, there must be a common sign for the excess the marginal tax over T^e (which is the sign of $1 - (\beta/M)$, or by (A.15), the sign of $-\mu(\varphi)$). But we see from (A.12) that a firm's average tax, $T(\varphi)/y(\varphi)$, must rise above the industry's average tax, T^e, because it is below this average tax at low outputs. This is part (c). The only way for this to happen is for the marginal tax to rise above T^e at some output level. Thus, this marginal tax must be everywhere greater than T^e. This is part (a).

Part (d) then holds because a firm's average tax is less than T^e at y^{**}, then rises above T^e, but eventually must fall because the marginal tax falls to T^e at y^h. Q.E.D.

Note that Proposition A does not imply that the marginal tax on output declines with output over the entire range of taxed outputs, as illustrated in Figure 3, but this property is consistent with the proposition. Whether there are intervals of output where the marginal tax is increasing will depend on the behavior of the term, $\frac{(1 - F(\varphi))}{f(\varphi)} c_{\varphi y}(y(\varphi), \varphi)$, in (A.17).

References

Auerbach, Alan and James R. Hines. 2002. "Taxation and Economic Efficiency." In A. Auerbach and M. Feldstein (eds.), *Handbook of Public Economics*, Volume 3, North-Holland.

Diamond, Peter and James Mirrlees. 1971. "Optimal Taxation and Public Production, Part I: Production Efficiency." *American Economic Review* 61 (March): 8-27.

Gordon, Roger H. and Wei Li. 2005. "Tax Structure in Developing Countries: Many Puzzles and a Possible Explanation." NBER Working Paper No. 11267, April.

Heller, Walter P. and Karl Shell. 1974. "On Optimal Taxation with Costly Administration." *American Economic Review Papers and Proceedings* 64 (May): 338-345.

Hopenhayn, Hugo A. 1992a. "Entry, Exit, and Firm Dynamics in Long Run Equilibrium." *Econometrica* 60(5) (September): 1127-1150.

Hopenhayn, Hugo A. 1992b. "Exit, Selection, and the Value of Firms." *Journal of Economic Dynamics and Control* 16(3-4): 621-653.

International Monetary Fund (with input from the staff of the Inter-American Development Bank, OECD, and the World Bank). 2007. *Taxation of Small and Medium Enterprises*. Background paper for the International Tax Dialogue Conference, Buenos Aires. October.

Keen, Michael and Jack Mintz. 2004. "The Optimal Threshold for a Value-Added Tax." *Journal of Public Economics* 88(3) (March): 559-576.

Kopczuk, Wojciech and Joel Slemrod. 2006. "Putting Firms into Optimal Tax Theory." *American Economic Review Papers and Proceedings* 96 (May): 130-134.

Melitz, Marc J. 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71(6): 1696-1725.

Mirrlees, James. 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies*, 38 (April): 175-208.

Slemrod, Joel. 2007a. "The Economics of Tax Remittance." Mimeo. University of Michigan.

Slemrod, Joel. 2007b. "Optimal Taxation with Tax Enforcement Externalities." In process. University of Michigan.

Wilson, John D. 1989. "On the Optimal Tax Base for Commodity Taxation. *American Economic Review*. 79(5) (December): 1196-1206.

Yitzhaki, Shlomo. 1979. "A Note on Optimal Taxation and Administrative Cost." *American Economic Review* 69 (June): 475-480.

Zee, Howell. 2005. "Simple Analytics of Setting the Optimal VAT Exemption Threshold." *De Economist* 153(4): 461-471,