Journal of Public Economics 95 (2011) 1036–1047

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Tax policy and the missing middle: Optimal tax remittance with firm-level administrative costs☆

Dhammika Dharmapala [a], Joel Slemrod [b,*], John Douglas Wilson [c]

[a] University of Illinois at Urbana-Champaign, 504 East Pennsylvania Avenue, Champaign, IL 61820, United States
[b] University of Michigan, 701 Tappan Street, Ann Arbor, MI 48109-1234, United States
[c] Michigan State University, Marshall-Adams Hall, East Lansing, MI 48824, United States

ARTICLE INFO

Article history:
Received 30 June 2009
Received in revised form 12 October 2010
Accepted 20 October 2010
Available online 3 December 2010

Keywords:
Taxation
Optimal taxation
Administrative costs

ABSTRACT

We analyze the optimal taxation of firms when the government faces fixed (per-firm) administrative costs of tax collection. The tax instruments at the government's disposal are a fixed (per-firm) fee and a linear tax on output. If all firms in an industry are taxed, we show that it is optimal to impose a positive fee to internalize administrative costs. The output taxes satisfy the inverse elasticity rule for taxed industries, but industries with sufficiently high administrative costs should be exempted from taxation. We also investigate the case where firms with outputs below a cutoff level can be exempted from taxation. It may be optimal to set the cutoff high enough to exempt a sizable number of firms, even though some firms reduce their outputs to the cutoff level, creating a "missing middle": small and large firms – but not those of intermediate size – exist. Thus, this common phenomenon in developing countries may result from optimal policies. The paper also presents a modified inverse-elasticity rule when output cutoffs are used, and it extends the analysis to include optimal nonlinear taxes on output.

1. Introduction

Some students of developing country economies have noted a phenomenon of the "missing middle", in which many small firms and a few large firms produce the bulk of value added. It forms part of a wider cluster of interrelated characteristics – including a large informal sector and regulatory barriers to entry into the formal sector – that have often been explained with reference to the "grabbing hand" associated with bureaucratic inefficiency and corruption (e.g. Friedman et al., 2000; Djankov et al., 2002). In the development literature, it has been suggested that the missing middle arises in part because taxes and regulations are enforced only among large, formal-sector firms (e.g. Rausch, 1991).[1] Notably, though, the development literature does not address under what conditions it is *optimal* for policy to treat differentially firms of different sizes, perhaps in a way that generates a missing middle. Nor, indeed, has the public finance literature focused on this issue, largely because its standard normative framework does not allow for heterogeneous firm size in a meaningful way.

In this paper, we provide a normative framework for analyzing policies that apply differentially to small and large firms, and demonstrate that under some conditions optimal policies will generate a missing middle — a range of output that no firm produces. We address tax policy, although some of the insights are also relevant to regulatory policy. Our basic argument is that the government should economize on administrative costs by exempting small firms from taxation, even though doing so causes intermediate-sized firms to reduce their outputs to tax-exempt levels, thereby creating a missing middle.[2] More broadly, a central aim of our paper is to construct a model of optimal tax policy that recognizes the central role that firms play in the remittance systems of all modern taxes, and the potential importance of treating firms differently according to their size, regardless of whether these taxes are levied in developing or developed countries.

☆ We are grateful for comments on an earlier draft received from the Editor (Emmanuel Saez), two anonymous referees, Alan Auerbach and attendees of seminar presentations at Hunter College, the London School of Economics, Queen's University, Stanford University, the University of California at Berkeley, University College London, the World Bank, the Office of Tax Policy Research conference on "The Role of Firms in Tax Systems," and the 2008 Centre for Business Taxation symposium at Oxford University.

* Corresponding author. Tel.: +1 734 936 3914; fax: +1 734 936 8716.
E-mail addresses: jslemrod@umich.edu, jpubeinvitations+slemrod@gmail.com (J. Slemrod).

[1] Tybout (2000) reviews the empirical evidence about this phenomenon.

[2] Our approach is consistent with Gordon and Li's (2009) basic contention that the policies pursued by developing countries can potentially be explained on normative grounds, rather than through political economy explanations based on the self-interest of non-benevolent policymakers.

0047-2727/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2010.10.013

The importance of firms 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.[3] However, dealing with *small* businesses is not generally cost-efficient, and many tax systems partially or entirely exempt small businesses from remittance responsibility.[4] Although special tax treatment of small firms might economize on compliance and administrative costs, 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, and more generally the design of the remittance of taxes, has not been closely addressed by the optimal tax literature.[5] 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. This is especially true in developing countries, where administrative constraints are often first-order issues in tax systems.

In addition, the famous Diamond and Mirrlees (1971) theorem on aggregate production efficiency posits 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 administrative costs and assumes that there are no untaxed profits, due either to constant returns to scale or a 100% tax on profit. Yitzhaki (1979) and Wilson (1989) investigate optimal commodity taxation when there is costly tax administration, but their assumption of constant returns to scale eliminates any role for firms. An earlier paper by Heller and Shell (1974) presents a general framework for analyzing an optimal system of commodity taxes, lump-sum taxes, and firm-specific licensing fees and 100% tax on profit when these tax instruments are costly to administer. In contrast, a major goal of the present paper is to investigate the optimal taxation of individual firms when firm-specific taxes are not available and therefore exemptions from taxation must be based on observed outputs. Keen and Mintz (2004) consider an output cutoff for exempting firms from a value-added tax in the presence of administrative and compliance costs. However, firms expect to earn untaxed profits in their model, in which case the Diamond–Mirrlees theorem no longer holds and, in general, different firms should be taxed at different rates, even without administrative costs.[6] As we next explain, in our model administrative costs are solely responsible for the production inefficiencies, and these costs require an expansion in the set of tax instruments.

In our main model, the three available policy instruments are a constant tax rate on output, a fixed per-firm fee, and an output cutoff, below which firms are not taxed. We show that when all firms in an industry are taxed, optimal policy involves the use of the fixed fee to internalize the social costs of tax administration.[7,8] In our model, each *industry* is characterized by constant returns to scale, because there is an effectively unlimited number of *ex ante* identical potential producers. We assume that firms “discover” heterogeneous productivities after entering the industry. 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. In particular, the Diamond and Mirrlees (1971) 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 cases in which it is optimal to exempt small firms from taxation. This creates production inefficiencies that are never part of an 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 the missing middle described above.

It is important to emphasize that our claim is a theoretical one – that a “missing middle” can potentially be generated by optimal tax policies under certain circumstances – rather than an empirical one (that observed “missing middles” correspond to existing tax thresholds). This caveat applies *a fortiori* to our results on the magnitude of the “missing middle”. It is also important to note, however, there is an emerging body of empirical evidence showing that firm size can be affected by various tax and regulatory thresholds. For instance, Onji (2009) analyzes the introduction of a value-added tax (VAT) in Japan in 1989. The new VAT system incorporated preferential treatment for small firms, with a cutoff for eligibility of 500 million yen in sales. Onji (2009) finds a clustering of firms just below this threshold following the reform. Labor market regulations also often vary by firm size. In Italy, firms with more than 15 employees face significantly more stringent employment protection regulations (and, in particular, higher firing costs). Using different empirical approaches, Garibaldi et al. (2004) and Schivardi and Torrini (2008) find significant effects of this threshold, involving slower firm growth and greater persistence close to the threshold.[9] This evidence provides support for the empirical importance for the distribution of firm size of tax and regulatory thresholds of the type analyzed in our model.

As a precursor to our model of heterogeneous firms, we first investigate the optimal system of output taxes and fixed fees in an economy where firms differ only across industries, not within

[3] The centrality of firms is illustrated by two recent studies that find that over 80% of all taxes are remitted by business in the U.S. and the U.K. — see Christensen et al. (2001) and Shaw et al. (2010). Anecdotal evidence suggests that collection of taxes from businesses is even more important in developing countries.

[4] For an excellent review of the sorts of special regimes that countries apply to small businesses, see International Tax Dialogue (2007). In many countries the exemption of small firms is *de facto*, due to lax enforcement. One implication of these policies is that the collection of taxes is highly concentrated among relatively large firms. A recent report asserts that the typical distribution of tax collections by firm size for African and Mid-Eastern countries features less than 1% of taxpayers remitting over 70% of revenues, and the report gives specific examples of highly concentrated patterns: in Argentina, 0.1% of enterprise taxpayers remit 49% of revenues; and in Kenya, 0.4% remit 61% (International Tax Dialogue (2007)). In contrast, most manufacturing *employment* in developing countries is in small firms (with less than 10 employees); see e.g. Tybout (2000). Gauthier and Gersovitz (1997) document the concentration of tax payments for Cameroon.

[5] 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. The importance of the remittance system is discussed in Slemrod (2008).

[6] This discussion refers to the second of two models presented by Keen and Mintz (2004). 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 allowing a per-firm fixed fee as a policy instrument.

[7] International Tax Dialogue (2007, p. 31) discusses the fixed per-firm fee, also known as a *patente* system, as an example of a presumptive tax regime. The fixed per-firm fee could also be motivated by the entry fees and registration costs imposed by governments on firms, as measured and analyzed by Djankov et al. (2002). Auriol and Warlters (2005) also argue that entry barriers may be optimal in some circumstances, but the reason is very different: the entry barriers generate rents for incumbents that the government can tax.

[8] If, instead of administrative costs borne in the first instance by the taxing authority, 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.

[9] In addition, there is widespread support for the more general notion that the distribution of firm size is influenced by taxation and regulation in developed as well as developing countries. Pagano and Schivardi (2003) document significant differences in firm size distributions across European countries, and attribute these in part to tax and regulatory policies. Henrekson and Johansson (1999) find that Sweden has a particularly small share of medium-sized firms (with 10–199 employees), and explain this with reference to tax and regulatory policies penalizing small firms.

industries. In particular, in Section 2 we characterize the optimal system of output taxes and fixed fees, and we describe a rule to determine when administrative costs are sufficiently high to justify the exemption of an industry from taxation. Section 3 develops the heterogeneous-firm model on which the rest of the paper relies. In Section 4, we show that under certain conditions, cutoffs are part of the optimal tax structure. In Section 5, we develop a modified inverse-elasticity rule for the average net (of administrative costs) taxes on different goods. In Section 6, 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. Section 7 concludes.

2. The structure of firm-remitted taxes across homogeneous industries

We begin by focusing on differences across industries rather than on differences among firms within an industry. To do so, 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 the derivative, $c_i'(y_i)$, is positive at all y_i, and c_{ei} is treated as an entry cost. When we later introduce heterogeneous firms, we will assume an additional fixed production cost incurred by firms after entering and choosing to produce. In either case, the entry cost and increasing marginal cost imply U-shaped average cost curves.

The tax instruments consist of a constant marginal tax rate on output, t_i, and a fixed fee, b_i. The fixed fee will turn out to be a critical component of the tax system when there are per-firm 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 profit function, defined ignoring the fixed entry cost:

$$\pi_i(p_i, b_i) = p_i y_i(p_i) - c_i(y_i(p_i)) - b_i. \tag{1}$$

Free entry guarantees that p_i will settle where $\pi_i\,(p_i,\,b_i)$ equals the entry cost.

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, which are zero in equilibrium. Labor serves as the *numeraire*, so the costs described above are measured in units of labor. So that we can work with demand functions for each good that depend only on the good's own price, we assume that the direct utility function is separable and quasi-linear in labor: $U_I(X_I)+\ldots+U_I(X_I)-L$. 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, $\Sigma_i\, v_i(q_i)$, where the wage is suppressed because it is fixed at one. In the presence of a unit tax t_i on good i, the consumer price q_i equals p_i+t_i. The number of firms producing good i, M_i, is determined by the requirement that demand equals supply: $X_i(q_i)=M_i\;y_i\;(p_i)$. By Walras' Law, these equilibrium conditions and the budget constraints for each agent (including the government) imply that the labor market clears.

In this paper we focus on the implications for optimal tax policy of a fixed per-firm cost of collecting taxes.[10] 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 the firms' output levels.[11] Although to simplify the model we assume that administrative costs are fixed and hence independent of the size of the firm, the qualitative results apply as long as there is a relatively large fixed component to these 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, it is convenient to take as control variables the consumer and producer prices, rather than the tax rates.[12] It is also convenient to treat the M_i's as control variables.

The Lagrangian describing the government's problem is as follows:

$$\begin{aligned} L = {} & \sum_i v_i(q_i) + \sum_i \lambda_i(X_i(q_i) - M_i y_i(p_i)) \\ & + \lambda_B\left(\sum_i (M_i((q_i - p_i) y_i(p_i) + b_i - A_i)) - E\right) + \sum_i \beta_i(\pi_i(p_i, b_i) - c_{ei}). \end{aligned} \tag{2}$$

The government seeks to maximize the utility of the representative consumer, subject to three types of constraints. 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 applies to the government budget constraint, where E is an exogenous revenue requirement. Finally, each multiplier β_i pertains to the requirement that profits equal zero in industry i.

We first show that the fixed fee should equal the per-firm administrative costs, leaving the output tax to finance expenditure needs. (All proofs appear in the Appendix.)

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

The value of b_i can be thought of as a kind of Pigouvian tax: a taxed firm generates a social cost in the form of administrative costs, and the fee should internalize this cost.[13] 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 addressed beginning in the next section, assuming all firms are taxed. Note that, *ceteris paribus*, the higher is A_i (and therefore the optimal value of b_i), the larger is the optimal output of a firm in industry i, and the lower is the optimal number of such firms. This is true because a higher b_i requires a higher p_i to maintain the zero-profit condition, which in turn implies a higher y_i; for given q_i and thus total output $X_i(q_i)$, this in turn implies a smaller M_i. This zero-profit condition implies that each firm produces at the bottom of its average cost curve, inclusive of b_i. Because Proposition 1 tells us that $b_i=A_i$, this minimization of each firm's average cost implies that its average production plus administrative costs is minimized.

[10] The model does not allow firms to split up or combine for tax purposes.

[11] See Slemrod (2006) for a review of the empirical evidence showing that the *compliance* costs of business taxes as a fraction of firm size fall with firm size, suggesting a significant fixed aspect of this part of tax collection costs.

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

[13] It is not a classic Pigouvian tax, as it addresses a fiscal externality that would not be present if no tax revenue were required.

With the fee addressing the external nature of the administrative costs, we should expect the output tax to satisfy the usual inverse-elasticity rule for an optimal tax system.[14] This turns out to be the case. Let α denote the marginal utility of income, and define the price elasticity of demand, measured positively, as follows:

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

We then have:

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

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

Proposition 2 asserts that, as b_i addresses the externality, any remaining net revenue requirement ought to be collected as would otherwise be optimal, in this case by following the well-known inverse elasticity rule.

In the case of heterogeneous firms considered in the next section, the set of producing firms and the set of taxed firms may be differentiated by establishing a size threshold for being subject to tax. But under the current assumption of homogeneous firms, producing firms can be distinguished from taxed firms only by exempting entire industries from taxation. The exemption of an industry might be optimal because, although administrative costs do not affect the optimal structure of output taxes on taxed goods (by Proposition 2), they may affect the optimal 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. The preceding argument 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 should respond to the higher A_i by raising the fixed fee b_i, and lowering b_j. If, however, A_i gets sufficiently high, then the government should reason that, 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 altogether in order to obtain the savings in administrative costs. Indeed, one can derive a sufficient condition for industry exemption that obtains if the ratio of administrative costs to output tax revenue exceeds a lower bound that depends on the marginal cost of funds.[15]

This argument has interesting parallels to the work of Yitzhaki (1979) and Wilson (1989), both of which consider the optimal taxation of a continuum of goods that enter a representative consumer's utility function symmetrically.[16] In these models, taxing any good incurs an administrative cost that varies across goods, which suggests that taxing fewer goods is better than taxing many goods. But expanding the tax base reduces the standard deadweight loss from taxation and, at the optimal number of taxed goods (tax base breadth), the marginal administrative cost equals the marginal saving in deadweight loss.

In the Yitzhaki and Wilson models, however, the reason that administrative costs vary across goods is exogenous. Nor could the source of variation be related to administrative costs at the firm level, as these models adopt the standard assumption that all firms exhibit constant returns to scale, rendering the size of firms indeterminate. In the present model, U-shaped average cost curves limit the equilibrium size of firms. The heterogeneity of production technologies introduces heterogeneity across sectors in the cost 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 inter-industry differences in administrative costs and obtain a tradeoff between these costs and the deadweight cost of taxation that is similar to the tradeoff studied by Yitzhaki and Wilson: industries with many small firms are likely to exhibit relatively high administrative costs in tax collection, increasing the likelihood that they should be exempted from taxation.

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 administrative costs have an important fixed per-firm component, it follows that it might be optimal to exempt small firms in an industry from taxation. As previously noted, the explicit or *de facto* exemption of small firms is a widespread phenomenon in tax systems. Thus it is especially important to have a theoretical framework that will enable the rigorous analysis of the optimal structure of such policies.

The treatment of firm heterogeneity in the model we develop is inspired by Hopenhayn (1992a,b) and Melitz (2003)**.** The papers by Hopenhayn examine stationary equilibria for a stochastic model in which firm-level productivity shocks follow a Markov process, generating a pattern of entry and exit by competitive firms. Melitz examines the steady-state equilibrium for a model in which monopolistically competitive firms learn their productivities immediately after entering the industry. Our model is a static version of the Melitz model, extended to include taxes, except that we follow Hopenhayn by assuming that firms behave competitively.[17]

In the model developed below, firms that are *ex ante* identical choose whether to enter an industry (taking into account expected profits and taxes), competing away expected profits to zero. Following Melitz (2003), however, incurring the cost of entry enables a firm to ascertain its productivity, and it chooses its output accordingly. Thus, the model endogenously generates variation in firm size, and also allows firms to endogenously adjust both their entry decisions and output choices in response to tax policies (including any exemptions or cutoffs). These features of the model constitute significant advantages relative to alternative approaches to modeling firm heterogeneity. For instance, the traditional “dominant firm” model that has been extensively used in the industrial organization literature typically involves a single large firm that exercises price leadership, surrounded by a “competitive fringe” of small price-taking firms. While this model also entails heterogeneity in firm size, this heterogeneity is generated by the exogenously-imposed assumption that one of the firms in the industry is “dominant” in the sense of choosing its price first (see, e.g., Kydland (1979, p. 358)). Thus, it is unclear how the structure of firm size in such a model would respond to variations in tax policy.[18]

[14] The inverse-elasticity result applies because of our assumption that the utility function is separable and quasi-linear in labor. See Auerbach and Hines (2002) for a careful exposition of this model. To derive Eq. (4), we use Roy's identity, which implies that $\alpha = -v_i'(q_i)/X_i$.

[15] The marginal cost of funds is equal to $MCPF = \frac{1}{1-\frac{t_i}{q_i}\varepsilon_i^X}$. The sufficient condition is derived in Dharmapala et al. (2009).

[16] See also Slemrod and Kopczuk (2002), which adds distributional concerns to the choice of how broad the tax base should be.

[17] An explicit steady-state analysis is a straightforward extension of our model, assuming that the goal of tax policy is to maximize steady-state welfare.

[18] In addition, an earlier literature on the effects of taxation and regulation with an informal sector allows for firm heterogeneity (e.g. Fortin et al., 1997), but typically also involves *ex ante* exogenous differences among firms. Moreover, this literature does not derive optimal tax rules, as we do in this paper.

We first describe the behavior of firms, and then turn to the government's optimal tax problem. Because we are interested in how firms within a given industry should be taxed, at first 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 sub-optimization 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 optimal vector of consumer prices is then investigated in Section 6. In particular, 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 incur the same fixed cost to enter the industry, denoted c_e, but now allow that the convex variable cost function differs across firms: $c(y,\varphi)$ for a type-φ firm, where φ is an *ex ante* unknown productivity parameter that takes on values over an interval, $[\varphi^l, \varphi^h]$, with a density function that is strictly positive at each value within this range.[19] The value of φ cannot be discovered (by the entrepreneur) 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. By assuming a continuum of firms, we ensure that industry output is non-random, although (*ex ante*) any single firm's output is random.

Each firm chooses its output only after incurring the fixed cost, c_e. We assume that a higher value of φ decreases marginal costs: $\frac{\partial c(y,\varphi')}{\partial y} < \frac{\partial c(y,\varphi'')}{\partial y}$ for $\varphi' > \varphi''$ and all y. Thus, a firm's chosen output rises with φ. In addition, we assume that $c(y,\varphi)$ is bounded from below by a positive number at all positive y, reflecting a fixed cost incurred if any positive output is produced (in addition to the entry cost c_e), which we assume for simplicity to be constant across firms.

For a firm facing unit output price p and fixed fee b, profits, calculated ignoring the fixed entry cost, are given by

$$\pi(p,b,\varphi) = py(p,\varphi) - c(y(p,\varphi),\varphi) - b, \tag{5}$$

where the supply function, $y(p, \varphi)$, is obtained by maximizing profits. These profits are increasing in φ.

Once a firm has entered the industry, it incurs no additional costs if it exits the industry before producing (i.e., $c(y, \varphi)$ discontinuously drops to zero at $y = 0$). Thus, it will exit if the φ it draws is too small to yield non-negative profits; that is, if $\pi(p, b, \varphi) < 0$. Although exit entails no costs, the entry cost c_e is sunk and cannot be recovered. For the subsequent analysis, we assume that there always exist some firms whose productivity is low enough that they exit.[20]

Now consider introducing a tax-threshold level of output, y^*, such that firms with output equal to or less than y^* are not taxed and thus receive the consumer price q per unit of output. Firms with output above y^* pay the fixed fee b and are taxed at the rate t on (all) their output, in which case they receive the producer price p $(=q-t)$ per unit of output. Thus, we investigate the optimality of tax schedules that exhibit a discontinuous jump in tax liability as output rises above y^*.[21] The assumption here is that the government is unable to observe productivities directly, and must therefore base its tax on the observed output, which we assume can be observed by the tax authority without cost.[22]

If y^* is sufficiently low, then no firm will be willing to produce untaxed output, because a firm cannot cover its fixed costs if its output is too low. As y^* increases, however, eventually some firms producing above y^* will be tempted to lower their output to y^* to escape taxation. Higher levels of y^* can induce some firms to choose outputs below y^*. But there will still be firms that produce at y^*, but would produce above y^* if the tax break at y^* were not available.

Fig. 1 illustrates the incentives facing two such firms, which differ in their productivities. The optimally-taxed outputs for these high- and low-productivity firms are located where their marginal cost curves, MC^h and MC^l, equal the producer price p. The high-productivity firm incurs a loss in producers' surplus of area $I+II+III$ from the drop in output from y_h to y^*, but this drop is offset by the elimination of the tax burden once y^* is reached. The low-productivity firm incurs a smaller loss in producers' surplus, given by area I, because its marginal cost curve is higher by assumption. Thus, it too reduces output to y^*. There is then bunching at output y^* of firms with productivities within some interval. Moreover, this bunching eliminates the production of outputs between y^* and some higher output, giving us the "missing middle" proposition discussed in the Introduction:

Proposition 3. (*The "missing middle"*). *Under an optimal linear tax system with a cutoff y^* that induces some firms not to pay taxes, there exists a $y^{**} > y^*$ such that firms with sufficiently high productivities produce outputs above y^{**}, but no firms produce an output between y^* and y^{**}.*

Thus, the economy contains small firms and large firms, but is missing firms with intermediate levels of output. Firms that would be producing these outputs instead reduce their outputs to y^* to eliminate their tax burdens. The lowest taxed output that firms are willing to produce, y^{**} in Proposition 3, leaves the marginal firm indifferent between y^* and y^{**}:

$$qy^* - c(y^*,\varphi) = py^{**} - c(y^{**},\varphi) - b \equiv \pi(p,b,\varphi). \tag{6}$$

Solving this equality for φ gives us the lowest productivity possessed by firms producing taxable output, defined as a function of prices, y^*, and the fixed fee b: $\varphi^{**}(q, p, b, y^*)$. With this notation, the total output of taxed firms is

$$Y^T = M\left(\int_{\varphi^{**}(q,p,b,y^*)}^{\phi^h} y(p,\varphi)f(\varphi)d\varphi\right), \tag{7}$$

where M is the number of firms entering the industry. This output and the total output of untaxed firms are non-random, because there is a continuum of firms. Unlike the previous model, M is no longer the number of firms that actually produce in the industry, because firms exit if their productivities are too low.

Consider now the determination of the untaxed output. Let $\varphi^m(q, y^*)$ denote the lowest productivity of active untaxed firms. If y^* is not too high, there will be no firms producing below y^*. Then this lowest productivity will be determined by the zero-profit requirement, $qy^* - c(y^*, \varphi) = 0$, in which case it declines with q and y^*, noting that higher y^* reduces the inefficient reduction in output required to achieve tax-exempt status. But at higher values of y^*, some firms maximize profits at untaxed outputs below y^*, in which case small changes in y^* have no effect on the value of $\varphi^m(q, y^*)$. Instead, this minimum productivity is determined by the

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

[20] A firm that exits thus knows its productivity. This knowledge, however, does not affect entry and exit decisions in any way, as firms are assumed not to have the opportunity to reenter (and because firms with sufficiently low productivities that exit will have no incentive to reenter in any event).

[21] The case in which there is a discontinuous jump in tax liability at the threshold is not necessarily unrealistic. Indeed, thresholds for VAT registration typically operate in this manner, although some governments mitigate the discontinuity by applying a lower rate over some range; see the discussion in Keen and Mintz (2004, pp. 560–562).

[22] In reality, observing output is not completely without cost, although we are confident that observing output is less costly than observing, for example, profits. A more complete model of the process would consider an enforcement system that audits output and presumably deters understatement. We believe that adding this feature to the model would not fundamentally alter the paper's results.

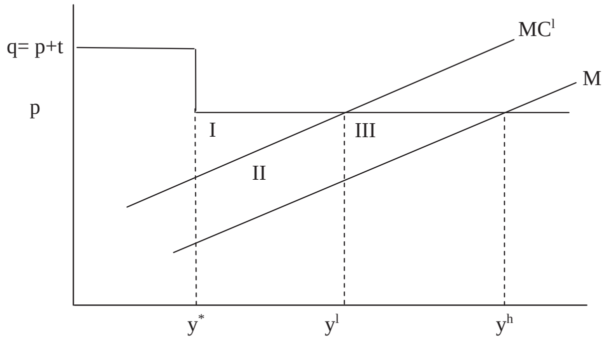

Fig. 1. Bunching under Linear Taxation.

requirement that maximized profits in the absence of taxes equal zero: $\pi(q, 0, \varphi) = 0$. In addition, the minimum productivity among bunched firms, denoted $\varphi^*(q, y^*)$, is the productivity at which $y(q, \varphi) = y^*$, whereas $\varphi^*(q, y^*) = \varphi^m(q, y^*)$ if no firms produce below y^*. In either case, total untaxed output may be written

$$Y^U = M\left(\int_{\varphi^m(q,y^*)}^{\varphi^*(q,y^*)} y(q,\varphi)f(\varphi)d\varphi + \int_{\varphi^*(q,y^*)}^{\varphi^{**}(q,p,b,y^*)} y^*f(\varphi)d\varphi\right), \tag{8}$$

where the first integral disappears if y^* is reduced to the point where no firm desires to produce below y^*. M is determined by the requirement that total supply equals demand:

$$Y^U + Y^T = X(q). \tag{9}$$

To fully demonstrate a missing middle, we must set up the government's optimal tax problem and characterize cases where it is indeed optimal for some firms to produce untaxed output. As discussed above, we first fix q and find the tax system that maximizes government revenue. The Lagrangian for this problem is:

$$\begin{aligned} L = {} & M\left[\int_{\varphi^{**}(q,p,b,y^*)}^{\varphi^h} ((q-p)y(p,\varphi) + b - A)f(\varphi)d\varphi\right] \\ & + \lambda\left(X(q) - M\left[\int_{\varphi^m(q,y^*)}^{\varphi^*(q,y^*)} y(q,\varphi)f(\varphi)d\varphi + \int_{\varphi^*(q,y^*)}^{\varphi^{**}(q,p,b,y^*)} y^*f(\varphi)d\varphi + \int_{\varphi^{**}(q,p,b,y^*)}^{\phi^h} y(p,\varphi)f(\varphi)d\varphi\right]\right) \\ & + \beta\left(\int_{\varphi^m(q,y^*)}^{\varphi^*(q,y^*)} \pi(q,0,\varphi)f(\varphi)d\varphi + \int_{\varphi^*(q,y^*)}^{\varphi^{**}(q,p,b,y^*)} (qy^* - c(y^*,\varphi))f(\varphi)d\varphi \right. \\ & \left. + \int_{\varphi^{**}(q,p,b,y^*)}^{\phi^h} \pi(p,b,\varphi)f(\varphi)d\varphi - c_e\right). \end{aligned} \tag{10}$$

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 that do not exit: 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 zero-profit constraint has an important implication: because firms' sales revenue $qX(q)$ must equal production costs plus tax payments, the maximization of tax payments net of administrative costs is equivalent to the minimization of total production costs plus administrative costs. To depict this minimization problem, we can write total production costs as a function of the number of taxed firms, $C(M^T)$, where the tax parameters are optimally chosen to achieve this number M^T; then the optimal M^T minimizes $C(M^T) + AM^T$. This is an unconstrained maximization problem because the $C(M^T)$ reflects not only the adjustment to the producer price p to achieve zero profits (for the given consumer price q), but also the entry of firms to equate supply with demand. If a cutoff is desirable, it should be set so that raising it enough to reduce the number of taxed firms by a unit causes production costs to rise by an amount equal to the saving in administration costs: $-dC/dM^T = A$.

Note that production plus administrative costs are minimized only in a second-best sense, given the limitations on the available tax instruments. The second-best inefficiencies inherent in a cutoff rule will be studied below. If we could identify firm productivities and implement a productivity cutoff rather than an output cutoff, then $C + AM^T$ could be lower.

This minimization problem suggests a definition of deadweight loss from taxing any particular good that generates firm-level administrative costs: it is the excess of $C + AM^T$ over the minimum production costs, where the latter would be achieved in this competitive economy by setting y^* equal to zero and taxing all firms at the same rate. An important insight is that it is not generally optimal to simply minimize production costs when collecting taxes incurs administrative costs.

Returning to the revenue-maximization problem, 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:

$$\lambda = \frac{R}{X} \equiv T^e. \tag{11}$$

The basic idea is that if there is an exogenous unit increase in output $X(q)$, then the number of firms in the industry can be increased to satisfy this output expansion, and the resulting increase in tax revenue, T^e, measures the social gain.

In the absence of a cutoff, T^e would equal the output tax rate, t, because the fixed fee b would equal administrative costs A, in its role as a Pigouvian tax. We show in an additional online Appendix, however, that the optimal t exceeds T^e when firms do take advantage of a cutoff.[23] Intuitively, the higher output tax, coupled with a lower fixed fee, makes relatively low levels of taxed output more attractive to firms, discouraging some of them from bunching at y^*. In other words, the fee is now no longer solely serving the role of a Pigouvian tax, but is also being used to control bunching. Given these conflicting considerations, we are unable to sign the difference between b and A when there is an output cutoff, although the excess of t over T^e limits the extent to which b can exceed A. Identifying this sign is not, however, needed for our subsequent results.

4. Missing middle or isolated bottom?

The missing middle identified in Proposition 3 suggests an economy with concentrations of small and large firms, but with no intermediate-sized firms. But another possibility is that the level of the output cutoff used to exempt firms from taxation is so low that only a small number of firms take advantage of it, generating a size distribution of firms with intermediate- and large-sized firms and a few much smaller untaxed firms, representing an “isolated bottom” of the size distribution. The next proposition rules out at least certain forms of this size distribution under an optimal output cutoff.

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.*

[23] The online Appendix is available at: http://www.bus.umich.edu/otpr/papers.htm. See in particular Proposition A.1, which also shows that $b = A$ with heterogeneous firms and no output cutoff. Proposition A.2 states that $t > T^e$ under a sufficient condition that seems unlikely to be violated.

The basic idea here is that if only a few firms can gain by choosing to produce untaxed output, then they cannot benefit much. Start with a cutoff y^* that is so low that no firms produce y^*, but then increase y^* until those active firms with the lowest output, y_1, are now indifferent between continuing to produce y_1 and instead reducing their outputs to y^*. 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. In terms of Fig. 1, if MC^l is the marginal cost curve for one of these firms, then the profit loss from reducing output to y^*, equal to area I, approximately equals the benefit of the tax reduction once y^* is reached, $b+ty^*$. 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.[24] As previously explained, welfare cannot be maximized if revenue is not maximized, given the consumer price q.

As the cutoff level is increased, firms initially producing slightly above y_1 now choose to reduce their outputs to y^*; and firms that were indifferent between y^* and y_1 now obtain higher profits at the higher y^*, since their production disadvantage from producing at y^* declines as y^* rises.[25] Thus, the cutoff starts to raise expected profits for the industry, enabling the government to increase its taxes on firms above the cutoff without causing the consumer price to rise. Whether the higher taxes more than offset the loss in revenue from the exit of more firms from the tax base will depend on their tax payments relative to their administrative costs. If their tax payments do not exceed administrative costs by much, then exempting them from taxation will have little effect on the government budget, so the higher taxes on firms above the cutoff will result in a budget surplus, which can be used to raise welfare. 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.*

Thus, the importance of administrative costs relative to tax collections is the critical determinant of whether there should be an output cutoff. 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 (but positive), however, then this second effect will be unimportant, 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 sufficient number of firms to produce untaxed output. Consequently, Proposition 5 lends further support for the theoretical optimality of the missing middle. In particular, a welfare-improving output cutoff induces some firms to opt out of the tax system by lowering their tax rates, so that there are many small untaxed firms as well as some large taxed firms, but there is a missing middle of intermediate-sized firms. Although Proposition 5 places an upper bound on values of T^e under which a welfare-improving output cutoff must exist, this upper bound is not necessarily small. However, it is important to bear in mind the caveat that the optimality of the missing middle is a theoretical result, and Propositions 4 and 5 provide only limited guidance as to its quantitative magnitude.

5. A modified inverse-elasticity rule

When there are no untaxed firms producing in any industry, the standard inverse- elasticity rule derived in Proposition 2 holds. In that case, only the demand elasticity matters because raising the consumer price alone has no impact on the producer price received by firms, the level of which is fixed by the requirement that expected profits equal zero. With output cutoffs, in contrast, this rule no longer applies. Untaxed firms receive the consumer price, so supply elasticities matter. Thus, to state a new rule, we must first define the supply elasticity of the untaxed output of good i:

$$\varepsilon_i^U = \frac{\partial Y_i^U}{\partial q_i}\frac{q_i}{Y_i^U} > 0. \tag{12}$$

For this definition, we hold fixed the *number* of firms producing the untaxed output and consider only the marginal impact of the price they receive (i.e., the consumer price) on their output.

A change in q_i, and any accompanying changes in the other tax variables, will also generally change the share of firms that produce untaxed output, denoted F_i^{**}. We define this share elasticity as follows:

$$\varepsilon_i^F = \frac{dF_i^{**}}{dq_i}\frac{q_i}{F_i^{**}} > 0. \tag{13}$$

The change in this share directly affects the net revenue obtained from taxing good i. Letting R_i denote this net revenue, we define the revenue elasticity,

$$\varepsilon_i^R = \frac{dR_i}{dF_i^{**}}\frac{F_i^{**}}{R_i}. \tag{14}$$

The additional online Appendix demonstrates that this revenue elasticity is negative (Proposition A.2).

These three elasticities all enter our modified inverse-elasticity rule. Again let α denote the consumer's marginal utility of income, and λ_B denote the marginal value of government revenue. With this notation, we now state the new rule as follows:

Proposition 6. *Assume that t_i and b_i are optimal for each good i, given exogenous (optimal or not) output cutoffs. Then the average net tax rate for each taxed good satisfies the following modified inverse-elasticity rule:*[26]

$$\frac{T_i^e}{q_i} = \frac{1-\frac{\alpha}{\lambda_B}}{\varepsilon_i^X + \frac{Y_i^U}{X_i}\varepsilon_i^U - \varepsilon_i^R\varepsilon_i^F} \tag{15}$$

This is a rule for the average net tax on output as a percentage of the consumer price — an average *ad valorem* net tax. Intuitively, it is the net tax rate that matters, not the gross rate, because a rise in output not only generates a social gain in the form of additional gross tax revenue, but also a social cost in the form of additional

[24] With p and b initially set to maximize tax revenue in the absence of a cutoff (subject to the requirement that profits equal zero, given q), the envelope theorem tells us that the welfare effect from adjusting y^* to induce a marginal number of firms to produce untaxed output does not depend on whether p and b also change by marginal amounts (but changes in p and b will affect the lowest level of y^* at which some firms are willing to produce untaxed output).

[25] No firm will choose to produce below y^* until y^* is increased to a sufficiently high level.

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

administrative costs. Observing that $t_i > T_i^e$ (Proposition A.2 in the additional online Appendix), this rule places lower bounds on the output tax rates.

The modified inverse-elasticity rule tells us that not only should, *ceteris paribus*, taxes be low on goods with high demand elasticities, but, other things equal, they should also be low on goods with high supply elasticities for untaxed output. In contrast, supply elasticities only matter in the Ramsey model when the standard assumption of constant returns to scale is replaced with decreasing returns, implying positive profits. In the current model, the assumption of free entry implies constant returns to scale at the industry level, and zero expected profits. But taxing output at a higher rate distorts not only demand decisions, but it 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 net revenue elasticity, which is shown in the additional online Appendix to be negative (Proposition A.2): the movement of a marginal firm from taxed to untaxed output lowers revenue. Thus, goods with high share elasticities should, *ceteris paribus*, be taxed at relatively low net rates.

Note, finally, that a good with a relatively high share of output that is untaxed, given by $\frac{Y_i^U}{X_i}$, should have a relatively low average tax, all else equal. The reason is that the supply elasticity for untaxed output becomes more important in the modified inverse-elasticity rule as the untaxed-output share rises.

Now consider the case where the government uses a cutoff for some industries but not others. Without a cutoff, only demand elasticities enter the inverse-elasticity rule. With a cutoff, the supply-related elasticities also enter, and they all contribute to a lower tax rate. Thus, we have shown:

Proposition 7. *Assume that t_i and b_i are optimal for each good i, given exogenous (optimal or not) output cutoffs. If taxed goods i and j have the same demand elasticity, but the government uses a cutoff rule only for i, then the optimal average net tax on output is lower for i than for j: $\frac{T_i^e}{q_i} < \frac{T_j^e}{q_j}$.*

Presumably, a cutoff rule is more likely to be used in industries with relatively high administrative costs. If this is the case, Proposition 7 tells us that the industries with the high administrative costs tend to have the lower average net taxes. In other words, any additional gross tax payments are not fully covering the higher administrative costs. The intuition is that, although a cutoff rule lowers the total cost of taxing a sector's firms, it increases the marginal cost of so doing because a higher tax rate causes both inefficient supply and demand responses.

Total administrative costs will tend to be relatively large for industries whose technologies dictate that they will consist of relatively small firms.[27] 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 7 also suggests that low taxes should be levied on small firms, net of these administrative costs. Of course, their gross taxes would be high enough to cover these administrative costs.

Note also that the preceding analysis assumes a tax that is linear (apart from the nonlinearity associated with the cutoff rule, when it is used). Thus, it does not shed light on the relationship between firm size and the tax rate, except to the extent that smaller firms are exempted altogether from the tax. The next section introduces more general nonlinear tax systems.

6. Nonlinear tax systems

Much of the potential welfare gain of an output cutoff for taxation may be offset by its negative impact on firms' output decisions, as firms attempt to qualify for tax-exempt status by reducing their outputs. One way to counteract these inefficiencies would be to give tax breaks to firms with taxable outputs near the cutoff. Simply stated, reducing the tax liability on those firms that might be tempted to cut their outputs to the cutoff level might keep them from so doing. This could be achieved using a nonlinear tax system under which a tax function, $T(y)$, defined net of administrative costs, is chosen so that firms choosing a relatively low output pay a relatively low average net tax, $T(y)/y$, where, as before, "net" means after administrative costs are subtracted.

In this section, we demonstrate that the revenue elasticity defined by Eq. (14) equals zero under an optimal nonlinear tax system. In other words, the tax schedule is chosen so that the movement of another firm from taxable output to nontaxable output has no effect on net revenue. Although the firm no longer pays taxes, its reduction in output necessitates additional entry into the industry, to maintain the equality between supply and demand, and this entry offsets the revenue loss. We then discuss the implications of this zero revenue elasticity for our modified inverse elasticity rule. Finally, we describe possible shapes of the optimal nonlinear tax schedule.

We again focus on a single industry and consider the sub-optimization 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 we allow this aspect of the tax system's nonlinearity to be supplemented by marginal output taxes, dT/dy, that vary with output at $y > y^*$ rather than being set at a constant rate of t.

The formal treatment of the optimal nonlinear tax problem is in the additional online Appendix. Here we provide a heuristic analysis based on Fig. 2, which presents a possible tax schedule, beginning at y^*. A profit indifference curve over tax payments and output is drawn for a type-φ^{**} firm, which 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^{**}. Firms with productivities between φ^{**} and some lower value of φ bunch at y^*.

This bunching could be eliminated by replacing the discontinuity 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.

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 outputs from y^{**} to $y^{**} + dy^{**}$. Because 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

[27] Note that the size distribution of firms depends not only on the exogenous properties of technologies, but also on tax policy. In particular, since a cutoff rule causes some firms to eliminate their tax burdens by reducing their outputs to the cutoff level, it tends to reduce the average size of firms.

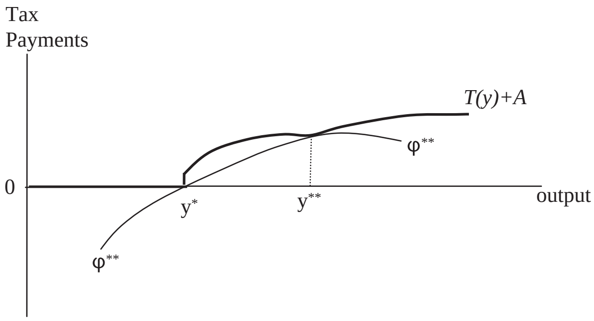

Fig. 2. Bunching under Nonlinear Taxation.

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 output per firm entering the industry. Thus, $dM = \frac{y^{**}-y^*}{y^e}$. Following our previous notation, let T^e be the average tax on output, calculated net of administrative costs, in which case the tax payment per firm is $T^e y^e$. Multiplying this amount by the change in M gives the total change in tax payments from this entry: $(y^{**} - y^*)\ T^e$. 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, the following condition must hold when the optimal number of firms is exempt from taxation:

$$T(y^{**}) = (y^{**} - y^*)T^e. \tag{16}$$

Condition (16) says that the nonlinear 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. As a result, we may conclude that the revenue elasticity defined by Eq. (14) equals zero.[28] An interesting aspect of this rule is that it contains no terms involving the behavioral responses of firms' outputs 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, when the government is restricted to a linear tax system, it can alter the number of taxed firms only by changing the common marginal output tax rate t, which causes *all* taxed firms to alter their chosen output supplies.

With a zero revenue elasticity, our modified inverse elasticity rule, given by Eq. (15), reduces to

$$\frac{T_i^e}{q_i} = \frac{1 - \frac{\alpha}{\lambda_B}}{\varepsilon_i^X + \frac{Y_i^U}{X_i}\varepsilon_i^U}. \tag{17}$$

Comparing Eq. (17) with Eq. (15), we see that if the government uses nonlinear taxes for some, but not all, goods, then, 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.

It is interesting to observe that Eq. (17) would also hold in the case where the government could exempt firms from taxation based on their productivities, rather than observations of output. There would be no “share elasticity” in this case, because firms would not be able to switch their tax status in response to a change in tax rates. For the tax systems studied in this article, the presumption is that a firm's productivity is not observable by the tax authority.

In our additional online Appendix, we derive tax schedules for firms above the cutoff y^{**} that possess the form shown in Fig. 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 Eq. (16). But the marginal tax on this output is high and declining, eventually falling below the average tax as the top output is approached.[29,30]

Although we have shown that allowing a nonlinear output tax can mitigate the distortions created in tax systems with an output cutoff rule, in practice this would clearly involve additional administrative costs. For this reason, this paper's attention to linear taxes is appropriate. 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 output levels with tax breaks for somewhat higher output levels.

7. Conclusions

To be relevant to a world (like ours) in which there are significant administrative costs to collecting taxes, 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 to the difficult problem of collecting taxes 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, and demonstrate that under some conditions the optimal tax system will generate a missing middle of firm sizes. To do so, we generalize the policy instruments available to the tax authority by allowing them to collect taxes from some, but not all, firms in an industry. In addition, we show that a fixed per-firm fee can be an important component of an optimal tax system, and that average net (of administrative costs) 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.

The models we have developed in this paper also demonstrate 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. This suggests that optimal industrial organization must be considered together with optimal tax policy.

Future modeling work might usefully extend some aspects of our models that are highly stylized. For example, the models in this paper presume that all goods are produced in a single stage of production and that each firm produces only one good. Thus, they cannot address the important production efficiency questions that arise in the analysis of cascading business turnover taxes (also known as gross receipts taxes). Nor can the present models address without some refinement an important issue that arises in the implementation of a value-added tax: the exemption of firms that sell to non-exempt firms (or that sell to exempt firms that sell to non-exempt firms) does not

[28] To obtain a zero revenue elasticity, combine Eq. (16) with (A6) in the Appendix. Establishing this result does not actually 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.

[29] Specifically, we show that $dT(y)/dy$ lies above T^e for $y^{**} \leq y < y(p, \varphi^h)$, and declines to T^e at the top output level $y(p, \varphi^h)$, if $(1 - F(\varphi))/f(\varphi)$ goes to zero as φ goes to φ^h. This condition clearly holds for a uniform distribution. As a result, the average tax, $T(y)/y$, reaches a maximum at some y between y^{**} and y^h, as shown in Fig. 3.

[30] The low rate on firms at y^{**} (where they just enter the tax net), combined with high and declining marginal tax rates thereafter, is somewhat analogous to results in the literature on optimal income taxation with a discrete labor force participation decision (e.g. Saez, 2002).

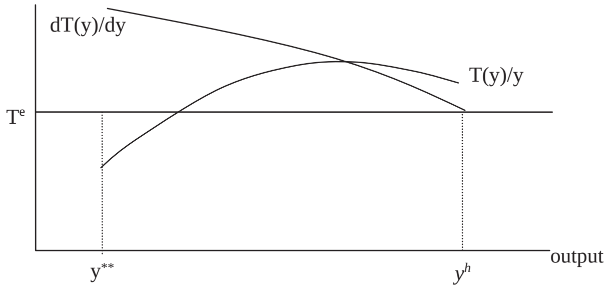

Fig. 3. Marginal and Average Tax Schedules under Nonlinear Taxation.

truly relieve these firms of some tax burden, unless it induces the formation of "chains" of exempt firms selling to each other and ultimately to final consumers.[31] It would be useful to allow the administrative cost to be itself subject to policy control, as stressed by Slemrod and Yitzhaki (2002).[32] Finally, taxation is not the only form of government intervention in the production process, and many other types of regulation (such as health and safety standards, or antidiscrimination laws) explicitly or implicitly exempt small firms, creating many of the same incentives addressed in this paper, although without the same direct implications for revenue.[33] Ideally, models of government policy toward small firms should address all aspects of taxation and regulation simultaneously.[34]

Appendix A

Proof of Proposition 1. Let $b_i > A_i$. Then lower 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 see that the 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 the output per firm. As a result, more firms must enter the industry to keep the 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, which corresponds to a welfare gain because the surplus in the government budget can be used, e.g., to lower q_i. By reversing the argument, we find that b_i cannot be less than A_i. Thus, $b_i = A_i$. Q.E.D.

Proof of Proposition 2. Differentiating the Lagrangian with respect to M_i, we obtain a first-order condition that implies

$$\frac{\lambda_i}{\lambda_B} = \frac{R_i}{X_i} = t_i, \tag{A1}$$

where R_i is the amount of revenue raised by taxes on firms in sector i, calculated net of administrative costs. The revenue per unit of output equals the output tax because the 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 Eq. (4). Q.E.D.

Proof of Proposition 3. This proposition clearly holds if $b + ty^* > 0$, because any firm producing output slightly above y^* would benefit from lowering output by the small amount needed to eliminate its tax burden. Thus, we need only show that this inequality holds. Suppose instead that $b + ty^* = 0$. Then, because marginal costs vary continuously with productivity, the presence of some firms at y^* implies that there will be other firms producing outputs slightly above y^* and paying almost no taxes. But the government would be incurring the administrative cost A to collect a negligible amount of taxes. By raising y^*, it would incur almost no revenue loss, while eliminating these administrative costs. Thus, the original y^* could not have been optimal. A similar argument shows that we cannot have $b + ty^* < 0$. Q.E.D.

Proof of Proposition 4. If no firms produce untaxed output, then we know that revenue is maximized by setting $b = A$. Given this equality, raise y^* to the highest point where there is no bunching ($\varphi^{**} = \varphi^*$). Then only type-φ^{**} firms are willing to produce untaxed output, and these firms are just indifferent to reducing their output to qualify for tax-exempt status. To induce a marginal number of firms to move to untaxed output, we may then either raise b or y^* or both; the welfare effect will be the same, given that b and p are initially optimized for the case of the untaxed firms. Let us differentiate the Lagrangian given by Eq. (10) for the revenue-maximization problem with respect to y^*:

$$\frac{\partial L}{\partial y^*} = T^e(y^{**} - y^*)\frac{\partial F^{**}}{\partial y^*} - ty^{**}\frac{\partial F^{**}}{\partial y^*} = -ty^*\frac{\partial F^{**}}{\partial y^*} < 0, \tag{A2}$$

where $T^e = t$ when $b = A$, use is made of Eq. (11), and $\partial F^{**}/\partial y^*$ is defined as the marginal rise in the share of firms that produce untaxed output. Thus, starting from the highest cutoff at which there are no untaxed firms, inducing some firms to choose untaxed output creates a first-order revenue loss for a fixed value of q, implying that welfare must fall. Q.E.D.

Proof of Proposition 5. Returning to the Lagrangian given by Eq. (10) for the revenue-maximization problem, set $b = A$ (as required by the extension of Proposition 1 to heterogeneous firms in the absence of a cutoff) and $t = 0$, in which case increasing y^* alone has no impact on government revenue net of administrative costs. Differentiating the Lagrangian with respect to y^* gives

$$\frac{\partial L}{\partial y^*} = \beta \int_{\varphi^*}^{\varphi^{**}} \Big(q - c_y(y^*, \varphi)\Big) f(\varphi)\, d\varphi, \tag{A3}$$

where use is made of Eq. (11). As y^* increases above the highest level where no firm strictly benefits from producing y^* ($\varphi^{**} = \varphi^*$), firms bunch at y^* ($\varphi^{**} > \varphi^*$) and further increases in y^* raise profits for these

[31] On the formation of VAT chains, see de Paula and Scheinkman (2007).

[32] For example, Gordon and Li (2009) argue that when businesses and individuals have regular dealings with financial institutions, the cost of monitoring their tax affairs falls, and policy should effectively subsidize these interactions.

[33] A reviewer of a previous version of this paper asked whether it might be optimal to have a "minimum production threshold" below which firms are not allowed to operate. The answer is "no" for the current model. Such a threshold would cause some firms to produce more than would otherwise be optimal. Thus, firms would be larger on average. With the consumer price determining the total demand, however, the total output would not change; fewer firms would enter the industry, offsetting the larger size of firms. Thus, the total revenue from the output tax would not change. Since it is still optimal to set $b = A$, any saving in administrative costs would be offset by lower revenue from the fixed fee. But the distortion to production would reduce average profits, requiring a fall in the output tax rate to maintain zero profits, given the chosen consumer price. Thus, tax revenue would decline, implying that the minimum production threshold could not have been beneficial.

[34] Krueger (2009) reports that, since 1950, in India firms employing less than 10 workers have been exempt from many regulations governing employment of workers, provision of pensions and other safety net items. In response, many larger firms do not register as required, and in other cases factories "have office doors with different names on each one in order to keep under the limit of ten!" (p. 24). The latter behavior would be an example of misrepresenting firm size, which is not addressed by the models of this paper. The reorganization of firms solely for tax purposes may, though, affect the measured size distribution of firms even if it does not affect actual production operations; see Sivadasan and Slemrod (2008) for an example in which a change in the tax treatment of partnerships in India distorted some income inequality measures based on firm-provided data. Note finally, as Tybout (2000) stresses, that in many countries tax and regulatory policies favor large over small companies.

bunched firms, because they sell at a price q ($=p$ because $t=0$) above their marginal costs at y^*. Thus, the expected profits available to a firm entering the industry rise. Eq. (A3) 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. Because 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.

Proof of Proposition 6. To derive the new rule, recall the previous sub-optimization 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 whether a cutoff rule is used. The optimal tax problem then consists of maximizing the indirect utility function, subject to the government budget constraint:

$$Max \sum\nolimits_i v_i(q_i) \;\; s.t. \;\; \sum\nolimits_i R_i(q_i) = E,$$

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

$$\frac{\alpha}{\lambda_B} X_i(q_i) = \frac{dR_i(q_i)}{dq_i}, \tag{A4}$$

where α is the consumer's marginal utility of income, and λ_B is the Lagrange multiplier on the government's budget constraint, or the marginal social value of government revenue. 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 Eq. (10) with respect to q and this change in b, and using the previously-derived equality, $\lambda = T^e$, gives

$$\frac{dR}{dq} = X + T^e \frac{dX}{dq} - T^{**} M \frac{dF^{**}}{dq} - T^e \left(\frac{\partial Y^U}{\partial q} - (y^{**} - y^*) M \frac{dF^{**}}{dq} \right), \tag{A5}$$

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.

Changing the share of firms that produce untaxed output will alter tax revenue, R_i, by

$$\frac{dR}{dF^{**}} = M(T^e(y^{**} - y^*) - T^{**}). \tag{A6}$$

In expression (A6), T^{**} is the net revenue obtained from a marginal taxed firm, which is now lost as some of these firms move from output y^{**} to untaxed output y^*, and T^e $(y^{**} - y^*)$ is the resulting rise in revenue from the resulting entry of additional firms to increase total output back to total demand to offset the lower per-firm output of the untaxed firms. In the text, we proved that this derivative is equal to zero under optimal nonlinear taxation. But, in the case of linear taxation, we show in the online Appendix that revenue falls. Using Eq. (A6), we may rewrite Eq. (A5) as

$$\frac{dR}{dq} = X + T^e \left(\frac{dX}{dq} - \frac{\partial Y^U}{\partial q} \right) + \frac{dR}{dF^{**}} \frac{dF^{**}}{dq}. \tag{A7}$$

Substituting this derivative into Eq. (A4) and expressing the result in elasticity form yields Eq. (15) in Proposition 6. Q.E.D.

Appendix B. Proofs of Additional Propositions and Formal Analysis of the Optimal Nonlinear Tax Problem

Supplementary data to this article can be found online at doi:10.1016/j.jpubeco.2010.10.013.

References

Auerbach, Alan, Hines Jr., James R., 2002. Taxation and economic efficiency. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, Volume 3. North Holland.

Auriol, Emmanuelle, Warlters, Michael, 2005. Taxation base in developing countries. Journal of Public Economics 89 (4), 625–646 (April).

Christensen, K., Cline, Robert, Neubig, Thomas, 2001. Total corporate taxes: hidden, above-the-line, and non-income taxes. National Tax Journal 54 (3), 495–506 (September).

De Paula, Áureo, Scheinkman, José, 2007. The informal sector. NBER Working Paper No. 13486. (October).

Dharmapala, Dhammika, Slemrod, Joel, Wilson, John D., 2009. Tax Policy and the Missing Middle: Optimal Tax Remittance with Firm-Level Administrative Costs. Working paper.

Diamond, Peter, Mirrlees, James, 1971. Optimal taxation and public production, part I: production efficiency. The American Economic Review 61 (1), 8–27 (March).

Djankov, Simeon, La Porta, Rafael, López de Silanes, Florencio, Shleifer, Andrei, 2002. The regulation of entry. Quarterly Journal of Economics 117 (1), 1–37 (February).

Fortin, Bernard, Marceau, Nicholas, Savard, Luc, 1997. Taxation, wage controls and the informal sector. Journal of Public Economics 66 (2), 293–312 (November).

Friedman, Eric, Johnson, Simon, Kauffmann, Daniel, Zoido-Lobaton, Pablo, 2000. Dodging the grabbing hand: the determinants of unofficial activity in 69 countries. Journal of Public Economics 76 (3), 459–493 (June).

Garibaldi, Pietro, Pacelli, Lia, Borgarello, Andrea, 2004. Employment protection legislation and the size of firms. Giornale degli Economisti e Annali di Economia 63, 33–68.

Gauthier, Bernard, Gersovitz, Mark, 1997. Revenue erosion through tax exemption and evasion in Cameroon, 1993. Journal of Public Economics 64 (3), 407–424 (June).

Gordon, Roger H., Li, Wei, 2009. Tax structure in developing countries: many puzzles and a possible explanation. Journal of Public Economics 93 (7-8), 855–866 (August).

Heller, Walter P., Shell, Karl, 1974. On optimal taxation with costly administration. American Economic Review Papers and Proceedings 64 (2), 338–345 (May).

Henrekson, Magnus, Johansson, Dan, 1999. Institutional effects on the evolution of the size distribution of firms. Small Business Economics 12, 11–23 (February).

Hopenhayn, Hugo A., 1992a. Entry, exit, and firm dynamics in long run equilibrium. Econometrica 60 (5), 1127–1150 (September).

Hopenhayn, Hugo A., 1992b. Exit, selection, and the value of firms. Journal of Economic Dynamics and Control 16 (3-4), 621–653 (July-October).

International Tax Dialogue (with input from the staff of the International Monetary Fund, 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, Mintz, Jack, 2004. The optimal threshold for a value-added tax. Journal of Public Economics 88 (3), 559–576 (March).

Krueger, Anne O., 2009. The missing middle. working paper no. 230. Indian Council for Research on International Economic Relations. January.

Kydland, Finn, 1979. A dynamic dominant firm model of industry structure. Scandinavian Journal of Economics 81 (3), 355–366.

Melitz, Marc J., 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71 (6), 1695–1725 (November).

Onji, Kazuki, 2009. The response of firms to eligibility thresholds: evidence from the Japanese value-added tax. Journal of Public Economics 93 (5-6), 766–775 (June).

Pagano, Patrizio, Schivardi, Fabiano, 2003. Firm size distribution and growth. Scandinavian Journal of Economics 105 (2), 255–274 (June).

Rausch, James E., 1991. Modeling the informal sector formally. Journal of Development Economics 35, 33–47 (January).

Saez, Emmanuel, 2002. Optimal income transfer programs: intensive versus extensive labor supply responses. Quarterly Journal of Economics 117 (3), 1039–1073 (August).

Schivardi, Fabiano, Torrini, Roberto, 2008. Identifying the effects of firing restrictions through size-contingent differences in regulation. Labour Economics 15 (3), 482–511 (June).

Shaw, Jonathan, Slemrod, Joel, Whiting, John, 2010. Administration and Compliance, in Institute of Fiscal Studies (Ed.), Dimensions of Tax Design: The Mirrlees Review, Oxford University Press.

Sivadasan, Jagadeesh, Slemrod, Joel, 2008. Tax law changes, income shifting, and measured wage inequality: evidence from India. Journal of Public Economics 92 (10-11), 2199–2224 (October).

Slemrod, Joel. 2006. The compliance cost of taxing business. Mimeo. University of Michigan.

Slemrod, Joel, 2008. Does it matter who writes the check to the government? The economics of tax remittance. National Tax Journal 61 (2), 251–275 (June).

Slemrod, Joel, Kopczuk, Wojciech, 2002. The optimal elasticity of taxable income. Journal of Public Economics 84 (1), 91–112 (April).

Slemrod, Joel, Yitzhaki, Shlomo, 2002. Tax avoidance, evasion, and administration. In: Auerbach, Alan, Feldstein, Martin (Eds.), Handbook of Public Economics. Elsevier, Amsterdam, London, and New York.

Tybout, James R., 2000. Manufacturing firms in developing countries: how well do they do, and why? Journal of Economic Literature 38 (1), 11–44 (March).

Wilson, John D., 1989. On the optimal tax base for commodity taxation. The American Economic Review 79 (5), 1196–1206 (December).

Yitzhaki, Shlomo, 1979. A note on optimal taxation and administrative cost. The American Economic Review 69, 475–480 (June).

Zee, Howell, 2005. Simple analytics of setting the optimal VAT exemption threshold. De Economist 153 (4), 461–471.