Production versus Revenue Efficiency with Limited Tax Capacity: Theory and Evidence from Pakistan

Michael Carlos Best

Stanford University

Anne Brockmeyer

World Bank

Henrik Jacobsen Kleven

London School of Economics

Johannes Spinnewijn

London School of Economics

Mazhar Waseem

University of Manchester

To fight evasion, many developing countries use production-inefficient tax policies. This includes minimum tax schemes whereby firms are taxed on either profits or turnover, depending on which tax liability is larger. Such schemes create nonstandard kink points, which allow for eliciting evasion responses to switches between profit and turnover taxes using a bunching approach. Using administrative data on corporations in Pakistan, we estimate that turnover taxes reduce evasion by up to

We thank Tony Atkinson, Michael Devereux, Jim Hines, Agnar Sandmo, Tim Schmidt-Eisenlohr, Jesse Shapiro, Monica Singhal, Joel Slemrod, and five anonymous referees for

Electronically published October 30, 2015
[*Journal of Political Economy*, 2015, vol. 123, no. 6]

60–70 percent of corporate income. Incorporating this in a calibrated optimal tax model, we find that switching from profit to turnover taxation increases revenue by 74 percent without reducing aggregate.

I. Introduction

A central result in public economics is that tax systems should maintain full production efficiency even in second-best environments (Diamond and Mirrlees 1971). This result permits taxes on consumption, wages, and profits but precludes taxes on intermediate inputs, turnover, and trade. The result has been very influential in the policy advice given to developing countries, but a key concern with such advice is that the underlying theoretical assumptions are ill-suited to settings with limited tax capacity. In particular, the result is based on an environment with perfect tax enforcement—zero tax evasion at zero administrative costs—which is clearly at odds with the situation in developing countries. Once we allow for tax evasion or informality, it may be desirable to deviate from production efficiency if this leads to less evasion and therefore greater revenue efficiency. While there is some theoretical work along these lines (e.g., Emran and Stiglitz 2005; Gordon and Li 2009), there is virtually no empirical evidence on the trade-off between production and revenue efficiency in the choice of tax instruments.[1]

To address this question empirically, we exploit a production inefficient tax policy commonly observed in developing countries. This is the imposition of minimum tax schemes according to which firms are taxed either on profits or on turnover (with a lower rate applying to turnover), depending on which tax liability is larger.[2] This policy has been motivated by the idea that the broader turnover base is harder to evade, an argu-

discussions and comments. Financial support from the Economic and Social Research Council grant ES/J012467/1 is gratefully acknowledged. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations or those of the executive directors of the World Bank or the governments they represent. Code is provided as supplementary material online.

[1] A few studies have taken a macro cross-country approach focusing on trade vs. domestic taxes (Baunsgaard and Keen 2010; Cage and Gadenne 2014).

[2] Such minimum tax schemes have been implemented in numerous developing countries, including Argentina, Bolivia, Cambodia, Cameroon, Chad, Colombia, Democratic Republic of the Congo, Ecuador, El Salvador, Equatorial Guinea, Gabon, Guatemala, Guinea, Honduras, India, Ivory Coast, Kenya, Laos, Madagascar, Malawi, Mauritania, Mexico, Morocco, Nigeria, Pakistan, Panama, Philippines, Puerto Rico, Republic of the Congo, Rwanda, Senegal, Taiwan, Tanzania, Trinidad and Tobago, and Tunisia (see Ernst & Young [2013] for a description). Most of these minimum tax schemes are based on turnover, but a few of them are based on alternative bases such as total assets or broader taxable income measures in between profits and turnover.

ment that seems intuitive but is so far untested. Crucially, these minimum tax schemes give rise to kink points in firms' choice sets: the tax rate and tax base jump discontinuously at a threshold profit rate (profits as a share of turnover), but tax liability is continuous at the threshold. We show that such kinks provide an ideal setting for estimating evasion responses to switches between profit and turnover taxes using a bunching approach, allowing us to evaluate the desirability of deviating from production efficiency to achieve greater compliance.

The basic empirical idea is that excess bunching at the minimum tax kink will be driven (mostly) by evasion or avoidance responses rather than by real production responses. To see this, consider the firm-level incentives under a turnover tax as compared to a pure profit tax. Because turnover is a much broader base than profits, minimum tax schemes in general involve turnover tax rates that are much smaller than profit tax rates. In our application to Pakistan, the turnover tax rate is 0.5 percent while the profit tax rate is 35 percent. The low turnover tax rate implies that this tax introduces only a small distortion to real production at the intensive margin, while a profit tax levied on true economic profits would be associated with a zero distortion of real production at the intensive margin. Hence, the change in real production incentives around the kink is small. On the other hand, because the tax bases are completely different on each side of the kink, there will be a large change in evasion incentives if those bases are associated with different evasion opportunities. Hence, if we see large bunching at the minimum tax kink, this is difficult to reconcile with real output responses under reasonable elasticity parameters and provides prima facie evidence of an evasion response to the switch between turnover and profit taxation.

In the paper we provide two important clarifications to the argument that bunching represents tax evasion. First, while the preceding paragraph compares a turnover tax to a nondistortionary tax on pure profits, the argument is robust to allowing for distortionary profit taxes. The real production incentives introduced by distortionary profit taxes do not contribute to bunching at the kink, but create movements away from the kink. Second, the absence of real responses to the minimum tax kink does not imply that the economywide production distortions of turnover taxation are small. Overall distortions may be substantial because of general equilibrium cascading effects (taxing the same item multiple times through the production chain; see, e.g., Keen 2013) and extensive margin responses (as total tax liability can be large because of the broadness of the turnover base; see, e.g., Auerbach, Devereux, and Simpson 2010). However, bunching at the kink captures only the partial equilibrium intensive margin response. Bunching will therefore not be driven by these other effects, enabling us to bound the extent of evasion.

Using administrative data containing the universe of corporate tax returns in Pakistan between 2006 and 2010, we find large and sharp bunching in reported profit rates around the kink point that separates the turnover tax and profit tax regimes. By exploiting variation in the minimum tax kink over time and across firms, we show that the bunch moves with changes in the location of the kink, that it increases in the size of the kink, and that it completely disappears during a temporary elimination of the kink. These findings provide compelling nonparametric evidence that firms respond to the minimum tax incentives in the way that we expect. The weakness of real incentives around the kink suggests that evasion is an important part of the story. We consider a competing hypothesis in which lazy reporting of costs by firms that fall within the turnover tax regime contribute to bunching. Such reporting errors would lead us to overstate deliberate evasion, but we present an empirical test showing that lazy reporting is in practice not a key confounder.

Using a simple model and a range of assumptions about the real output elasticity, we convert our bunching estimates into evasion estimates. We show that turnover taxes reduce evasion by up to 60–70 percent of corporate income compared to profit taxes. The evasion estimates are not very sensitive to the size of the real output elasticity, because the small change in real incentives around the kink implies that real responses contribute very little to bunching even under very large elasticities.

We use our empirical estimates to analyze the optimal choice of tax base and tax rate. We find that a switch from profit taxation to turnover taxation (at a much lower tax rate) can increase corporate tax revenues by 74 percent without decreasing aggregate after-tax profits (hence representing a welfare gain). The reason is that the loss of production efficiency is more than compensated for by the increase in revenue efficiency due to larger compliance. While these gains are based on a uniform turnover tax on all firms, we argue that heterogeneity in evasion may justify a minimum tax regime that limits turnover taxation to a subset of firms with low reported profit rates.

Our paper contributes to several literatures. First, we contribute to the recent bunching literature (Saez 2010; Chetty et al. 2011; Kleven and Waseem 2013) by developing an approach that exploits the simultaneous discontinuity in tax rate and tax base. Our approach has wider applicability than the specific minimum tax scheme considered here, including policies such as the Alternative Minimum Tax in the United States. Second, we add to an emerging empirical literature on public finance and development using administrative microdata (Kleven and Waseem 2013; Kumler, Verhoogen, and Frías 2013; Pomeranz 2013; Carrillo, Pomeranz, and Singhal 2015). Third, a theoretical literature has studied the implications of limited tax capacity for optimal taxation (Emran and Stiglitz 2005; Keen 2008; Gordon and Li 2009; Kleven, Kreiner, and Saez 2009;

Dharmapala, Slemrod, and Wilson 2011). While most of these papers study movements between formal and informal sectors and do not develop quantitative implications for policy, our paper studies corporate tax evasion at the intensive margin and derives simple expressions for optimal tax policy that depend on parameters that we estimate. Fourth, our paper develops a novel quasi-experimental methodology for the estimation of evasion, which can easily be replicated in other contexts as the tax variation needed is ubiquitous.[3] Finally, we contribute to the large literature studying responses by corporations to the tax code (see Auerbach [2002], Hassett and Hubbard [2002], and Auerbach et al. [2010] for surveys).

The paper is organized as follows. Section II presents our theoretical model, which is used in Section III to develop an empirical methodology based on minimum tax schemes. Section IV describes the context and data, and Section V presents our empirical results. Section VI numerically analyzes optimal policy, while Section VII presents conclusions.

II. Theoretical Model

This section develops a stylized model of the optimal taxation of firms in which firms decide how much to produce and what to declare for tax purposes. The analysis considers a government setting both the tax rate and the tax base in the presence of tax evasion. Our model incorporates the notion that a tax on output (turnover) is harder to evade than a tax on profits, the argument being that it is harder to evade a broader base and that it may be easier to fabricate costs than to conceal revenues.[4] When tax enforcement is perfect, the optimal tax system leaves the firm's production decision undistorted by taxing profits. When tax enforcement is imperfect, it becomes optimal to move toward a distortionary tax on output if this discourages tax evasion by firms. We first consider the trade-off between production efficiency and revenue efficiency (compliance) in a partial equilibrium analysis of final good production and then show how the analysis extends to a general equilibrium setting with intermediate good production. The stylized model allows us to identify sufficient statistics that capture this trade-off and guides our empirical strategy in the next section. Specifically, the partial equilibrium analysis depends on sufficient statistics that we estimate empirically, whereas the

[3] A vast literature has tried to estimate tax evasion using a variety of macroeconomic and microeconomic approaches (as surveyed by Andreoni, Erard, and Feinstein [1998], Slemrod and Yitzhaki [2002], and Slemrod and Weber [2012]). However, except for the rare occasions in which randomized tax audits are available (e.g., Slemrod, Blumenthal, and Christian 2001; Kleven et al. 2011), methodological limitations mean that the credibility and precision of these estimates are questionable.

[4] When the output price is normalized to 1, turnover and output are identical, and so we will use the terms "output tax" and "turnover tax" interchangeably throughout the paper.

general equilibrium analysis depends on some additional parameters that we do not estimate in this paper.

A. *Tax Policy in Partial Equilibrium*

A firm chooses how much output y to produce at a strictly convex and differentiable cost $c(y)$. The firm may misreport costs $\hat{c} \neq c(y)$ at a strictly convex and differentiable cost of misreporting $g(\hat{c} - c(y))$ with $g(0) = 0$.[5] The firm pays taxes $T(y,\ \hat{c}) = \tau(y - \mu\hat{c})$ depending on its output and declared costs. The tax liability is determined by the tax rate τ and a tax base parameter μ. The tax base parameter is the share of costs that can be deducted from a firm's revenues when determining the tax base. The tax base thus ranges from an output tax base ($\mu = 0$) to a pure profit tax base ($\mu = 1$). The ability to misreport costs captures the ease of evading profit taxes relative to evading output taxes.

The firm chooses y and $\hat{c}$ in order to maximize after-tax profits,

$$\Pi(y,\ \hat{c}) = (1 - \tau)y - c(y) + \tau\mu\hat{c} - g(\hat{c} - c(y)). \tag{1}$$

The actual after-tax profits $\Pi(\cdot)$ are in general different from reported after-tax profits $\hat{\Pi} \equiv (1 - \tau)y - \hat{c} + \tau\mu\hat{c}$. At the firm's optimum,

$$c'(y) = 1 - \tau\frac{1 - \mu}{1 - \tau\mu} \equiv 1 - \tau_E, \tag{2}$$

$$g'(\hat{c} - c(y)) \geq \tau\mu. \tag{3}$$

The output level is decreasing in the effective marginal tax rate τ_E. This rate represents the tax wedge between the social and private returns to output. For a pure profit tax base ($\mu = 1$), this wedge disappears and the output choice is efficient, regardless of the statutory tax rate. For an output tax base ($\mu = 0$), the effective tax rate equals the statutory tax rate. The impact of the statutory tax rate τ and the base parameter μ on the firm's output choice depends on the implied change in the effective tax rate τ_E with $\partial\tau_E/\partial\tau \geq 0$ and $\partial\tau_E/\partial\mu \leq 0$. The change from a high tax rate on a profit tax base to a lower tax rate on a broader output tax base will affect the firm's output choice only if it affects the effective tax rate τ_E.

The level of evasion is increasing in the base parameter μ and is thus higher for a profit tax base than for an output tax base. The level of evasion is also increasing in the tax rate τ. The latter result relies on the assumption that the cost of evasion $g(\cdot)$ depends on the difference between reported and true costs rather than on the difference between reported

[5] The modeling of evasion (or avoidance) based on a convex and deterministic cost function $g(\cdot)$ originates from Mayshar (1991) and Slemrod (2001).

and true tax liability (Allingham and Sandmo 1972; Yitzhaki 1974). The tax capacity of the government determines the cost of trying to evade taxes. When this evasion cost is sufficiently high ($g'(0) > \tau\mu$), the firm reports its true tax liability and the tax is perfectly enforced, regardless of whether profits or output is taxed.

The government sets the tax parameters τ, μ to maximize welfare subject to an exogenous revenue requirement R. In this stylized framework, this amounts to maximizing after-tax profits (corresponding to aggregate consumption by firm owners) subject to the revenue requirement. We assume that the private cost of evasion $g(\cdot)$ is also a social cost.[6] Hence, the welfare objective of the government can be written as

$$W = \Pi(y, \hat{c}) + \lambda[T(y, \hat{c}) - R], \tag{4}$$

where the firm's choices satisfy (2) and (3) and $\lambda \geq 1$ denotes the (endogenous) marginal cost of public funds. When there is no evasion, the government's problem is simple.

Lemma 1 (Production efficiency in partial equilibrium). With perfect tax enforcement (defined as $g'(0) > 1 \geq \tau\mu$), the optimal tax base is given by the firm's pure profit (i.e., $\mu = 1$).

Proof. For $\mu = 1$, we have $\tau_E = 0$ and hence $c'(y) = 1$, which ensures first-best output under any tax rate τ. The government sets $\tau = R/[y - c(y)]$ to satisfy its revenue constraint. QED

When we allow for evasion, the government's tax revenue can be decomposed into the revenue based on the true tax base and the forgone revenue due to misreporting the base,

$$\begin{aligned} T(y,\ \hat{c}) &= \tau \times \underbrace{(y - \mu\hat{c})}_{\text{reported base}} \\ &= \tau \times \Bigg\{ \underbrace{[y - \mu c(y)]}_{\text{true base}} - \underbrace{\mu[\hat{c} - c(y)]}_{\text{unreported base}} \Bigg\}. \end{aligned}$$

The government can raise revenue by increasing the tax rate ($\tau \uparrow$) or increasing the tax base ($\mu \downarrow$). Increases in both the tax rate and the tax base create a larger effective tax rate ($\tau_E \uparrow$) and thus decrease the firm's real

[6] The assumption that the private and social costs of evasion are the same is important for efficiency and optimal tax results (Slemrod 1995; Slemrod and Yitzhaki 2002; Chetty 2009). Examples of social evasion costs include productivity losses from operating in cash, not keeping accurate accounting books, and otherwise changing the production process to eliminate verifiable evidence. While including the evasion cost as a social cost is the natural starting point for developing countries where the revenue loss from evasion is a first-order social concern, it is conceptually straightforward to generalize this assumption. As we demonstrate in the online appendix, if the social cost of evasion is a fraction κ of the private cost, then the evasion term in the optimal tax rule that we derive (proposition 1 below) is simply scaled by the factor κ. Hence, the qualitative mechanisms that we emphasize (but of course not the quantitative importance) survive as long as $\kappa > 0$.

output level. The evasion effects, on the other hand, are not symmetric: while a larger tax rate increases the level of misreporting, a larger tax base decreases the level of misreporting. We may state the following proposition.

Proposition 1 (Production inefficiency in partial equilibrium). With imperfect tax enforcement (defined as $g'(0) = 0 \leq \tau\mu$), the optimal tax base is interior, that is, $\mu \in (0, 1)$. The optimal tax system satisfies

$$\frac{\tau}{1-\tau} \cdot \frac{\partial \tau_E}{\partial \tau}(\mu) = G(\mu) \cdot \frac{\varepsilon_{\hat{c}-c}}{\varepsilon_y}, \tag{5}$$

where

$$\varepsilon_{\hat{c}-c} \equiv \frac{\partial(\hat{c}-c)}{\partial \tau\mu} \frac{\tau\mu}{\hat{c}-c} \geq 0$$

is the elasticity of evasion with respect to $\tau\mu$,

$$\varepsilon_y \equiv \frac{\partial y}{\partial(1-\tau_E)} \frac{1-\tau_E}{y} \geq 0$$

is the elasticity of real output with respect to $1 - \tau_E$, and $G(\mu) \equiv [\hat{c} - c(y)]/\hat{\Pi} \geq 0$ is the level of evasion as a share of reported profits. We have

$$\frac{\partial \tau_E}{\partial \tau}(\mu) = \frac{1-\mu}{(1-\tau\mu)^2} \geq 0.$$

Proof. See the Appendix.

In the presence of evasion, it is always optimal to introduce at least some production inefficiency by setting $\mu < 1$. To understand the optimal tax rule (5), note that the left-hand side $[\tau/(1-\tau)] \cdot [\partial \tau_E(\mu)/\partial \tau]$ reflects the effective distortion of real production. This production distortion is equal to $\tau/(1-\tau)$ when $\mu = 0$, equal to zero when $\mu = 1$, and typically monotonically decreasing between those two extremes.[7] At the social optimum, the production wedge must be equal to the ratio between the evasion and output elasticities $\varepsilon_{\hat{c}-c}/\varepsilon_y$ scaled by the evasion rate $G(\mu)$. The evasion rate is equal to zero when $\mu = 0$ and is typically increasing in μ.

The formula highlights the trade-off between production efficiency (captured by the real output elasticity) and revenue efficiency (captured by the evasion elasticity) when setting the tax base μ. If the evasion elasticity is small relative to the real output elasticity ($\varepsilon_{\hat{c}-c}/\varepsilon_y \approx 0$), the production efficiency concern will be strong relative to the revenue efficiency

[7] Here we use that $(\partial \tau_E/\partial \tau)(0) = 1$, $(\partial \tau_E/\partial \tau)(1) = 0$, and that the cross derivative $\partial^2 \tau_E/\partial \tau \partial \mu$ is everywhere negative whenever $\tau \in [0, 1/(2-\mu)]$. The latter condition is satisfied for any tax rate below 50 percent, a very weak condition on a corporate income tax rate.

concern, and so it will be socially optimal to move close to a pure profit tax by setting $\mu \approx 1$ such that

$$\frac{\tau}{1-\tau}\cdot\frac{\partial \tau_E(\mu)}{\partial \tau} \approx 0.$$

Conversely, if the evasion elasticity is large relative to the real output elasticity, the revenue efficiency concern will be relatively strong, and this makes it optimal to move toward the output tax by lowering μ, thereby simultaneously decreasing the evasion rate $G(\mu)$ and increasing the production wedge until equation (5) is satisfied.[8] The former case is arguably the one that applies to a developed country context, whereas the latter case captures a developing country context. Our stylized framework thus highlights the starkly different policy recommendations in settings with strong versus weak tax capacity. Finally, note that equation (5) also identifies sufficient statistics for evaluating the optimal tax policy in this partial equilibrium framework, which we will study empirically in Section VI.[9]

B. *Tax Policy in General Equilibrium*

We now extend our stylized model to incorporate intermediate good production. We confirm the optimality of a profit tax in the absence of evasion, in line with the production efficiency theorem in Diamond and Mirrlees (1971), and generalize the optimal tax rule in the presence of tax evasion. The extension sheds light on two key general equilibrium effects of firm taxation and how they affect the optimal tax rule. First, moving away from a pure profit base causes cascading of tax distortions through the production chain, distorting the input mix and scale of downstream firms. Second, moving away from a pure profit base has an incidence effect, as price changes shift income between the final good sector and the intermediate good sector.

To see these effects, consider an economy with two firms operating in different sectors. Firm A produces an intermediate good y_A using labor l_A. Firm B produces a final good y_B using labor l_B and the intermediate

[8] The optimal tax rate τ changes endogenously as μ changes to satisfy the revenue constraint.

[9] Our decomposition into real output and evasion elasticities is not in contradiction with the sufficiency of taxable income elasticities for welfare analysis (Feldstein 1999). It is possible to rewrite eq. (5) in terms of the elasticities of taxable profits with respect to the tax rate τ and the tax base μ, respectively. If taxable profits are more responsive to an increase in the tax rate than to an increase in the tax base, this implies a relatively low efficiency cost associated with the tax base increase and therefore a low optimal μ. The presence of evasion, however, suggests an explanation for why these taxable profit responses may diverge as evasion is expected to respond in opposite directions to an increase in the tax rate ($\tau \uparrow$) and an increase in the tax base ($\mu \downarrow$). Our empirical methodology builds on this decomposition into real responses and evasion.

good y_A. Both firms can misreport their costs for tax purposes by incurring an additive evasion cost $g_i(\cdot)$ for firm $i = A, B$. Firm i's tax liability is $T(y_i, \hat{c}_i) = \tau(p_i y_i - \mu \hat{c}_i)$, where p_i is the price of the good it produces. We normalize the price of the final good to $p_B = 1$. We denote the wage rate by w and assume that labor is supplied perfectly elastically at this wage.

Firm A has a linear technology that converts labor into output one for one, $y_A = l_A$. This simplifying assumption implies that the equilibrium price of the intermediate good is simply determined by

$$w = p_A \frac{1-\tau}{1-\tau\mu} = p_A(1-\tau_E),$$

equalizing the marginal cost and marginal benefit of producing the intermediate good. The production technology for firm B is given by $y_B = F(l_B, y_A)$, and so firm B's input decisions satisfy

$$w = F'_{l_B} \cdot \frac{1-\tau}{1-\tau\mu} = F'_{l_B} \cdot (1-\tau_E),$$
$$w = F'_{y_A} \cdot (1-\tau_E)^2.$$

The latter condition equalizes the marginal cost and benefit of using the intermediate good to increase final good production at price $p_A = w/(1 - \tau_E)$. A positive effective tax rate not only distorts the scale of production but also distorts the input mix in the final goods sector away from the intermediate good. As a result the marginal rate of technical substitution $\text{MRTS}_{l_B, y_A} = F'_{l_B}/F'_{y_A} = 1 - \tau_E$ is distorted and $F'_{y_A} > F'_{l_B}$. The use of intermediate inputs for production is taxed twice: once at the intermediate production stage and once at the final production stage. This illustrates a new source of production inefficiency from a turnover tax that arises in general equilibrium—cascading through the production chain. As in the partial equilibrium analysis, the production distortions can be avoided by using a profit tax such that the effective tax rate is zero. We can state the following lemma.

LEMMA 2 (Production efficiency in general equilibrium). With perfect tax enforcement (defined as $g'(0) > 1 \geq \tau\mu$), the optimal tax base is given by the firm's pure profit (i.e., $\mu = 1$).

Proof. Suppose $\tau_E > 0$. If l_B decreases by Δ while l_A increases by Δ, then $y_A = l_A$ increases by Δ. Because $F'_{y_A} > F'_{l_B}$, this implies that y_B increases. Hence, total production has increased using the same amount of primary input, implying that $\tau_E > 0$ cannot be optimal. Under $\tau_E = 0$ ($\mu = 1$), the revenue requirement can be satisfied by appropriately selecting τ. QED

In the presence of imperfect tax enforcement, each firm $i = A, B$ will declare cost $\hat{c}_i$ such that

$$g_i'(\hat{c}_i - c_i) = \tau\mu.$$

An increase in the tax base ($\mu \downarrow$) discourages evasion by both the firm producing the final good and the firm producing the intermediate good. The optimal tax rule trades off production efficiency and revenue efficiency in this general equilibrium setting.

PROPOSITION 2 (Production inefficiency in general equilibrium). With imperfect tax enforcement (defined as $g'(0) = 0 \leq \tau\mu$), the optimal tax base is interior, that is, $\mu \in (0, 1)$. The optimal tax system satisfies

$$\frac{\tau}{1-\tau} \cdot \frac{\partial \tau_E}{\partial \tau}(\mu) \cdot \left\{ \frac{\beta[1 + \alpha(\mu)]}{1 + (1-\beta)\varepsilon_{p_A}} \right\} = G(\mu) \cdot \frac{\varepsilon_{\hat{c}-c}}{\varepsilon_y}, \tag{6}$$

where

$$\alpha(\mu) = \mathrm{MRTS}_{l_b, y_A} \Big/ \left[1 + \mathrm{MRTS}_{l_b, y_A} \left(\frac{\partial l_B}{\partial \tau_E} \Big/ \frac{\partial y_A}{\partial \tau_E} \right) \right]$$

is the production wedge caused by cascading, $\beta = y_B/(p_A y_A + y_B)$ is the share of the final good sector in total turnover, and $\varepsilon_{p_A} = (\partial p_A/\partial \tau_E)/(\tau_E/p_A)$ is the elasticity of the intermediate good's price with respect to the effective tax rate.

Proof. See the Appendix.

The optimal tax rule is the same as in the partial equilibrium case, but with the addition of the term in braces capturing cascading and incidence effects in the general equilibrium setting. The numerator captures the impact of cascading: $\alpha(\mu) \geq 0$ represents the effect of distorting the input mix of the final good sector when increasing τ_E and is larger the more substitutable labor and the intermediate good are in production, that is, the smaller $(\partial l_B/\partial \tau_E)/(\partial y_A/\partial \tau_E)$ is. Since cascading distorts production in the final good sector, its impact on the optimal policy rule is larger when the share of output coming from the final goods sector, β, is larger. The larger the cascading effect is, the more important production efficiency concerns are, and the narrower the optimal tax base (larger μ).

The denominator captures an incidence effect as an increase in the effective tax rate τ_E increases the price of the intermediate good, shifting income from the final goods sector to the intermediate good sector, reducing the importance of the efficiency of final goods production. The more responsive the intermediate good's price is (larger ε_{p_A}) and the larger the intermediate sector is (larger $1 - \beta$), the less important production efficiency concerns are, and the broader the optimal tax base (smaller μ).

Note that the production wedge still disappears when $\mu = 1$, and so formula (6) implies that a pure profit tax remains suboptimal in general equilibrium. The formula highlights precisely how the trade-off between production efficiency and revenue efficiency changes in general equilibrium. Since the cascading and incidence effects have offsetting impacts on the importance of production efficiency in the optimal tax rule, it is ambiguous whether the implementation of the partial equilibrium tax rule over- or underestimates the optimal tax base. We can state the following corollary.

Corollary 1 (Partial versus general equilibrium). Assuming that $(\partial\tau_E/\partial\tau)(\mu)$ is monotonically decreasing and $G(\mu)$ is monotonically increasing in μ, the partial equilibrium model implies a smaller μ (broader tax base) than the general equilibrium model iff $\beta[1+\alpha(\mu)]/[1+(1-\beta)\varepsilon_{p_A}] > 1$.

Hence, the partial equilibrium analysis is more likely to overstate the case for output taxation when the share of output in the final goods sector is large (β is large) and when labor and intermediate inputs are highly substitutable in the final goods sector ($\alpha \gg 1$).

C. *General Second Best with Many Tax Instruments*

The previous sections have analyzed a highly stylized setting in which the government can raise revenue only by taxing firms. While this is an oversimplification, it may not be unreasonable to focus on firm taxation. High enforcement and/or administrative costs force governments to rely heavily on taxes collected from firms rather than individuals and mean that some form of taxation of firms will always be present in low–fiscal capacity environments. Indeed, this is what is observed in countries with low fiscal capacity (Gordon and Li 2009; Besley and Persson 2013) and how theory suggests governments should optimally respond (Kopczuk and Slemrod 2006; Dharmapala et al. 2011).

More importantly, the qualitative prediction of our model—that optimal policy deviates from production efficiency in the presence of evasion—applies to a broad class of second-best settings with many tax instruments. Specifically, our results are related to a classic insight in the public finance literature that when at least one commodity cannot be taxed and pure profits are not taxed at 100 percent, some production inefficiency becomes desirable, even with otherwise unrestricted tax instruments (Stiglitz and Dasgupta 1971; Munk 1978, 1980). In our model with evasion, when $\mu = 1$ so that production is efficient, the profit tax does not correspond to a tax on true economic rents: unreported profits net of evasion costs, $(\hat{c} - c) - g(\hat{c} - c)$, go untaxed. In this case, it will always be desirable to deviate from production efficiency if this allows the government to extract some of the untaxed rents. Starting from $\mu = 1$, introducing a little bit of production inefficiency produces only a second-order welfare loss, while the benefit of indirectly taxing economic rents from evasion is first-order. This conceptual argument is unaffected by the existence of other instruments such as production-efficient consumption taxes. Naturally, a richer model featuring a richer set of instruments would have quantitative implications for how far the government would wish to deviate from production efficiency, but the qualitative conclusion that some production inefficiency is optimal in the presence of evasion would remain unchanged.

III. Empirical Methodology Using Minimum Tax Schemes

Using our theoretical framework, this section develops an empirical methodology that exploits a type of minimum tax scheme common to many developing countries, including Pakistan, which we consider in the empirical application below. Under this type of minimum tax scheme, if the profit tax liability of a firm falls below a certain threshold, the firm is taxed on an alternative, much broader tax base than profits. The alternative tax base is typically turnover (e.g., in Pakistan), and we focus on this case to be consistent with our empirical application. We show that such minimum tax schemes give rise to (nonstandard) kink points that produce firm bunching and that the bunching incentives vary greatly on the real production and compliance margins. We develop our approach within the simple partial equilibrium model of Section II.A, because general equilibrium cascading effects of turnover taxation do not generate bunching. We also consider the robustness of our approach to relaxing a number of simplifying assumptions.

A. Minimum Tax Kink and Bunching

Firms pay the maximum of a profit tax ($\mu = 1, \tau = \tau_\pi$) and an output tax ($\mu = 0, \tau = \tau_y$), where $\tau_y < \tau_\pi$. Tax liability is calculated on the basis of their turnover and reported costs in the following way:

$$T(y,\ \hat{c}) = \max\{\tau_\pi(y - \hat{c}),\ \tau_y y\}. \tag{7}$$

Firms thus switch between the profit tax and the output tax when

$$\tau_\pi(y - \hat{c}) = \tau_y y \Leftrightarrow \hat{\pi} \equiv \frac{y - \hat{c}}{y} = \frac{\tau_y}{\tau_\pi}. \tag{8}$$

This implies a fixed cutoff τ_y/τ_π for the reported profit rate $\hat{\pi}$ (reported profits as a share of turnover): if the profit rate is higher than this cutoff, firms pay the profit tax; otherwise they pay the output tax. As the reported profit rate crosses the cutoff, the tax rate and tax base change discontinuously, but the tax liability (7) is continuous. Hence, this is a kink (a discontinuous change in marginal tax incentives) as opposed to a notch (a discontinuous change in total tax liability), but a type of kink conceptually different from those explored in previous work (Saez 2010; Chetty et al. 2011) because of the joint change in tax rate and tax base. This joint change affects the incentives for real output and compliance differentially. The marginal return to real output $1 - \tau_E$ changes from 1 to $1 - \tau_y$ when switching from profit to output taxation, whereas the marginal return to tax evasion $\tau\mu$ changes from τ_π to 0. Hence, for firms whose reported profit rate falls below the cutoff τ_y/τ_π in the absence of the minimum tax, the introduction of the minimum tax reduces real output (loss of production efficiency) and increases compliance (gain in revenue efficiency).

Figure 1 illustrates how the minimum tax kink at τ_y/τ_π creates bunching in the distribution of reported profit rates. The dashed line represents the distribution of reported profit rates before the introduction of a minimum tax (i.e., under a profit tax). Assuming a smooth distribution

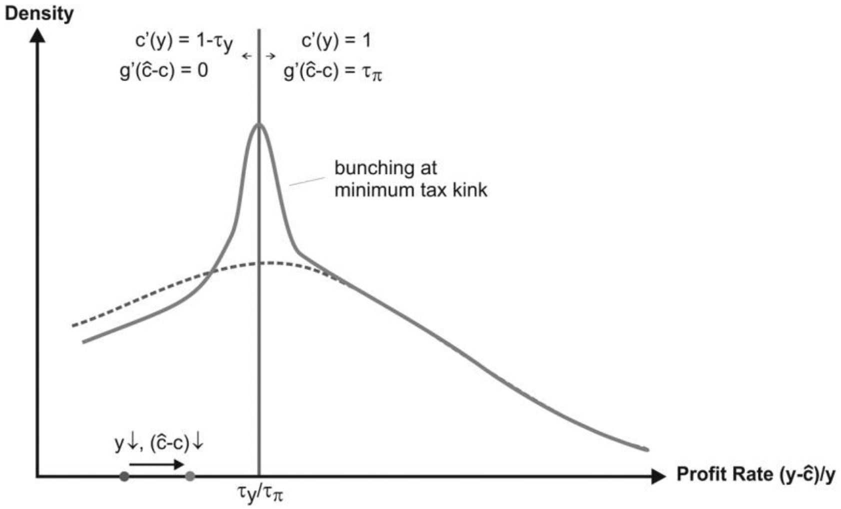

FIG. 1.—Bunching methodology using minimum tax schemes. The figure illustrates the implications of the introduction of a minimum tax on the observed density distribution of reported profit rates $\hat{\pi} = (y - \hat{c})/y$. The dashed line shows the smooth distribution of profit rates that would be observed in the absence of the minimum tax, while the solid line shows the distribution of profit rates that is observed in the presence of the minimum tax. As discussed in Section II, under the profit tax, firms' optimality conditions are given by $c'(y) = 1$ and $g'(\hat{c} - c) = \tau_\pi$. Firms whose optimal reported profit rate under the profit tax is smaller than τ_y/τ_π will adjust their production and reporting decisions in response to the introduction of the minimum tax to satisfy $c'(y) = 1 - \tau_y$ and $g'(\hat{c} - c) = 0$, causing them to decrease both output y and cost evasion $\hat{c} - c$. Both responses move their reported profit rate up toward the kink. Firms whose profit rate was close to the kink before introduction of the minimum tax pile up at the kink, which gives rise to an observed excess mass around the kink when accounting for optimization errors. Color version available as an online enhancement.

of firm profitability (through heterogeneity in either marginal production costs $c'(\cdot)$ or marginal evasion costs $g'(\cdot)$), this baseline distribution of profit rates is smooth and we denote it by $f_0(\hat{\pi})$. The introduction of a minimum tax (i.e., an output tax for $\hat{\pi} \leq \tau_y/\tau_\pi$) reduces the marginal return to real output from 1 to $1 - \tau_y$ and reduces the marginal return to evasion from τ_π to 0 for firms initially below the cutoff. These firms respond to the smaller real return by reducing output, which leads to an increase in their profit rates under decreasing returns to scale. They respond to the smaller evasion return by reducing tax evasion, which leads to a further increase in their reported profit rates. Both responses therefore create a right shift in the reported profit rate distribution below the cutoff (with no change above the cutoff) and produce excess bunching exactly at the cutoff. Allowing for optimization error (as in all bunching studies), there will be bunching around the cutoff rather than a mass point precisely at the cutoff, as illustrated in figure 1.[10]

Bunchers at the kink point τ_y/τ_π come from a continuous segment $[\tau_y/\tau_\pi - \Delta\hat{\pi}, \tau_y/\tau_\pi]$ of the baseline distribution $f_0(\hat{\pi})$ without the kink, where $\Delta\hat{\pi}$ denotes the profit rate response by the marginal bunching firm. Conceptually, it is the marginal buncher that reveals the underlying responsiveness to tax incentives as the inframarginal bunchers are restricted by their close proximity to the kink (see Saez 2010; Kleven and Waseem 2013). Assuming that bunching responses are local (such that $\Delta\hat{\pi}$ is small), the total amount of bunching is given by $B \approx \Delta\hat{\pi} \cdot f_0(\tau_y/\tau_\pi)$.[11] Hence, on the basis of estimates of excess bunching B and a counterfactual density at the kink, it is possible to infer the profit rate response $\Delta\hat{\pi}$ induced by the kink. This profit rate response can then be linked to the underlying responses on the real production and compliance margins. Totally differentiating $\hat{\pi} \equiv (y - \hat{c})/y$ and using the decomposition $d\hat{c} = d(\hat{c} - c) + dc$, we obtain

$$\Delta\hat{\pi} = \left[\frac{\hat{c}}{y} - c'(y)\right]\frac{dy}{y} - \frac{d(\hat{c} - c)}{y} \simeq \frac{\tau_y^2}{\tau_\pi}\varepsilon_y - \frac{d(\hat{c} - c)}{y}, \tag{9}$$

where we use that $c'(y) = 1$ and $\hat{c}/y = 1 - \hat{\pi} \simeq 1 - (\tau_y/\tau_\pi)$ in the vicinity of the cutoff. The output elasticity is defined as

[10] Note that real output reductions below the kink produce excess bunching only under decreasing returns to scale. In the case of constant (increasing) returns to scale, real output reductions below the kink would generate zero bunching (a hole) around the minimum tax kink. Hence, the possibility of nondecreasing returns to scale only strengthens our main conclusion below that bunching at minimum tax kinks must be driven primarily by evasion.

[11] This relationship uses that the density $f_0(\cdot)$ is roughly uniform in a small area $\Delta\hat{\pi}$ around the kink. If the density is strongly sloping around the kink or the bunching area $\Delta\hat{\pi}$ is not small, the relationship can be generalized to account for the slope in the underlying counterfactual density (see Kleven and Waseem 2013).

$$\varepsilon_y \equiv \frac{dy/y}{d(1-\tau_E)/(1-\tau_E)},$$

and we use that $d(1-\tau_E)/(1-\tau_E) = -\tau_y$ when crossing the kink.[12]

The bunching response $\Delta\hat{\pi}$ thus depends on both the real output response and the evasion response, but in very different ways. Consider first the case without evasion so that the profit rate response is directly proportional to the real output elasticity, $\Delta\hat{\pi} \simeq (\tau_y^2/\tau_\pi)\varepsilon_y$. In this case, large bunching (large $\Delta\hat{\pi}$) would translate into an extremely large output elasticity. This follows from the observation that τ_y^2/τ_π will in general be a tiny number, because output tax rates are always very small as a result of the broadness of the output base (e.g., τ_y is at most 1 percent in the case of Pakistan). The intuition for this result is that the combined changes in tax base μ and tax rate τ offset each other to create a very small change in the real return to output $1-\tau_E$, which makes the minimum tax kink a very small intervention in a model without evasion. Hence, the presence of large bunching around minimum tax kinks (which is what we find empirically) cannot be reconciled with believable real output elasticities in a model without tax evasion and therefore represents prima facie evidence of evasion.[13]

Once we allow for evasion in the model, it becomes possible to reconcile large bunching with believable output elasticities as the evasion response on the right-hand side of equation (9) closes the gap. While we cannot separately estimate real output and evasion responses using only one minimum tax kink, equation (9) allows for a bounding exercise on the evasion response under different assumptions about ε_y. Because of the smallness of the factor τ_y^2/τ_π, the estimated evasion response will be insensitive to ε_y. Furthermore, if, in addition to the presence of a minimum tax scheme, there is exogenous variation in the output tax rate τ_y

[12] The above characterization assumes homogeneous responsiveness across firms, implying that there exists a single marginal buncher that reveals $\Delta\hat{\pi}$ (inframarginal bunchers respond by less, but they would have been willing to respond by the same). This simplifies the exposition, but as shown by Saez (2010) and Kleven and Waseem (2013), it is possible to allow for heterogeneity in responsiveness, in which case bunching identifies the average responsiveness around the kink. Specifically, if we denote the underlying driver of heterogeneity by x, there exists a marginal buncher of type x that responds by $\Delta\hat{\pi}(x)$, and bunching identifies the average response across all x, $E_x[\Delta\hat{\pi}(x)]$. Equation (9) then splits the profit rate response into the production and compliance margins for an average firm responding by $\Delta\hat{\pi} = E_x[\Delta\hat{\pi}(x)]$.

[13] In theory, the real output elasticity could be very large if the production technology is close to constant returns to scale (the elasticity goes to infinity as we converge to constant returns to scale). However, near-constant returns to scale imply near-constant profit rates even under large output responses and therefore no output-driven bunching. In other words, real output elasticities are large precisely in situations in which output-driven bunching at the minimum tax kink must be small, and so the observation of large bunching cannot be credibly explained by real responses under near-constant returns to scale.

applying to this scheme (giving us more than one observation of $\Delta\hat{\pi}$ for the same values of the output elasticity ε_y and the evasion response $d(\hat{c} - c)$ under the exogeneity assumption), it may be possible to separately estimate the real and evasion responses. Variation in the profit tax rate τ_π is not useful for separately estimating output and evasion responses, because the profit tax rate directly affects the evasion response $d(\hat{c} - c)$ to the minimum tax kink (and so does not give us additional observations of $\Delta\hat{\pi}$ for the same values of $d(\hat{c} - c)$).

B. Robustness

The kink implied by the minimum tax scheme changes the incentives for production and evasion differentially. The analysis above shows that when combining a pure profit tax with a small turnover tax, the change in real incentives at the kink is minor, implying that substantial bunching provides evidence for evasion. This section shows that the key insight—that bunching at the minimum tax kink reflects mostly evasion—is robust to a number of generalizations.

Distortionary Profit Tax

The assumption that the profit tax corresponds to a tax on pure economic rent is very strong and stands in sharp contrast to a large body of literature analyzing the real distortions created by actual corporate income taxes (e.g., Hassett and Hubbard 2002; Auerbach et al. 2010). However, relaxing this assumption only strengthens our conclusion that observed bunching must be driven overwhelmingly by evasion responses. Other things equal, the introduction of real distortions in the profit tax regime implies that when firms move from profit to turnover taxation, real incentives will deteriorate by less or potentially improve. This additional effect by itself implies that the minimum tax scheme improves real incentives below the kink, so that firms respond by increasing their output. An increase in output reduces a firm's true profit rate ($\Delta\pi \leq 0$) under nonincreasing returns to scale and thus moves it away from the kink. Rewriting equation (9) as follows,

$$\Delta\hat{\pi} = \Delta\pi - d\left(\frac{\hat{c} - c}{y}\right), \tag{10}$$

we see clearly that in the case of a distortionary profit tax (implying $\Delta\pi \leq 0$ other things equal), real responses cannot be responsible for bunching at the minimum tax kink ($\Delta\hat{\pi} > 0$). We conclude that if the effective marginal tax rate under the profit tax were positive, our estimate of the evasion response based on the decomposition in (9) would provide a lower bound.

Output Evasion

We have so far emphasized cost evasion as the reason for the differential ease of evasion under profit versus turnover taxation, but this is not crucial for our empirical or conceptual results. Here we extend our model to allow for output evasion (reporting output $\hat{y}$ below true output y) in addition to cost evasion (reporting costs $\hat{c}$ above true costs c). Firm profits are given by

$$\Pi = y - c(y) - \tau(\hat{y} - \mu\hat{c}) - g(\hat{c} - c(y),\ y - \hat{y}),$$

where $g(\cdot)$ is now the total evasion cost on both margins. The analog of equation (9) becomes

$$\begin{aligned} \Delta\hat{\pi} &= \left[\frac{\hat{c}}{\hat{y}} - c'(y)\right]\frac{dy}{\hat{y}} - \frac{d(\hat{c} - c)}{\hat{y}} - \frac{\hat{c}}{\hat{y}}\frac{d(y - \hat{y})}{\hat{y}} \qquad (11) \\ &\simeq \frac{\tau_y^2}{\tau_\pi}\varepsilon_y\frac{y}{\hat{y}} - \frac{d(\hat{c} - c)}{\hat{y}} - \left(1 - \frac{\tau_y}{\tau_\pi}\right)\frac{d(y - \hat{y})}{\hat{y}}, \end{aligned}$$

decomposing the bunching response $\Delta\hat{\pi}$ into a real output response in the first term and the two evasion responses in the second and third terms. Three properties of this expression are worth noting. First, the model preserves the feature that the real response at the kink will be small as it is scaled by τ_y^2/τ_π. Second, the presence of bunching ($\Delta\hat{\pi} > 0$) corresponds to evasion reductions on either the cost margin ($-d(\hat{c} - c)/\hat{y} > 0$) or the output margin ($-d(y - \hat{y})/\hat{y} > 0$) when turnover taxation is introduced. Third, since in the third term we have $1 - \tau_y/\tau_\pi \approx 1$, bunching identifies approximately the aggregate evasion reduction on the two margins when switching from profit to turnover taxation. Hence, if there is less evasion on one margin (costs) but more evasion on the other margin (turnover) under turnover taxation, bunching captures the net effect of the two.

In both the empirical analysis and the calibration exercise in Sections V and VI, we show that our results are essentially unchanged when considering the more general model with output evasion.

Filing Costs (Lazy Reporting)

If firms face filing costs, bunching may be driven not only by tax evasion due to *cost overreporting* under the profit tax but also by filing errors due to *cost underreporting* under the turnover tax. To show this, we consider a model with fixed filing costs per item reported. That is, filing an item is costly because of the work and documentation that this requires, but the filing cost does not vary with the amount reported. Under this assumption, the turnover tax regime makes firms underreport the number of cost items, but not the amount per item reported. Such filing responses may create bunching at the minimum tax kink and therefore represent a potential confounder for our evasion estimates in Section V. We analyze this "lazy reporting" hypothesis in the empirical section.

To see this formally, consider a firm that incurs production costs as a continuum of items $j \in [0, 1]$ such that $c(y) = \int_0^1 c(y, j)dj$. For simplicity assume that the cost items are identical so that $c(y, j) = c(y)$ for all j. Including an item on the firm's tax return is costly, requiring the firm to incur a fixed filing cost $f(j)$, and we adopt the convention that cost items are ordered so that $f(j) > f(i)$ whenever $j > i$.[14] The firm can also overdeclare its costs in the categories it chooses to report, $\hat{c} > c$, at a convex cost $g(\hat{c} - c, j)$. For simplicity we assume that this cost is the same for all categories so that $g(\hat{c} - c, j) = g(\hat{c} - c)$ for $j \in [0, 1]$. If a firm produces output y and reports $\hat{c}$ in its first $N \leq 1$ cost categories, its profits are

$$\Pi(y, \hat{c}, N) = (1 - \tau)y - c(y) + N[\tau\mu\hat{c} - g(\hat{c} - c)] - \int_0^N f(r)dr.$$

In this case our empirical analysis will be based on observing declared profit rates $\hat{\pi} = (y - N\hat{c})/y$, and the analog of equation (9) becomes

$$\begin{aligned} \Delta\hat{\pi} &= \left[\frac{N\hat{c}}{y} - Nc'(y)\right]\frac{dy}{y} - \frac{Nd(\hat{c} - c)}{y} - \frac{N\hat{c}}{y}\frac{dN}{N} \qquad (12) \\ &= \left[\frac{\tau_y^2}{\tau_\pi} - \frac{\tau_y(1 - N)}{1 - N\tau_\pi}\right]\varepsilon_y - \frac{Nd(\hat{c} - c)}{y} - \left(1 - \frac{\tau_y}{\tau_\pi}\right)\frac{dN}{N}, \end{aligned}$$

where the final equality uses the firm's optimality condition that $c'(y) = (1 - \tau_\pi)/(1 - N\tau_\pi)$ and the fact that at the kink $N\hat{c}/y = 1 - \tau_y/\tau_\pi$.

Equation (12) decomposes the bunching response $\Delta\hat{\pi}$ into a real response in the first term, a cost overreporting response in the second term, and a response coming from a change in the number of cost items reported, dN/N, in the final term. Equation (12) also shows that in the presence of filing costs, when we use equation (9) to infer evasion responses from our estimated bunching response, we will be conflating evasion responses with filing responses coming from changes in the number of cost items reported, dN/N. To address this threat to our identification, in Section V.C we construct empirical measures of N and show that there is no evidence that N responds to the tax kink, providing reassurance that our estimates are capturing evasion responses rather than filing responses.

Pricing Power

The model can also be extended to incorporate pricing power by firms. In this case, firm profits are given by

[14] Realistically, firms also derive gains from keeping accurate tax records, and we should therefore think of $f(j)$ as net filing costs, which can be negative for some items.

$$\Pi = (1 - \tau)\rho(y)y - c(y) + \tau\mu\hat{c} - g(\hat{c} - c(y)),$$

where $\rho(y)$ is the price the firm receives, which depends negatively on output y. In this model, the analog of equation (9) is

$$\begin{aligned}\Delta\hat{\pi} &= \left[\frac{\hat{c}}{y}(1 - \sigma) - c'(y)\right]\frac{dy}{\rho(y)y} - \frac{d(\hat{c} - c)}{\rho(y)y} \\ &\simeq (1 - \sigma)\frac{\tau_y^2}{\tau_\pi}\varepsilon_y - \frac{d(\hat{c} - c)}{\rho(y)y},\end{aligned} \tag{13}$$

where

$$\sigma \equiv -\frac{\partial\rho(y)}{\partial y}\frac{y}{\rho(y)} > 0$$

is the price elasticity the firm faces and the second equality follows by using $\hat{c}/\rho(y)y = 1 - \tau_y/\tau_\pi$ and $c'(y) = \rho(y)(1 - \sigma)$ at the kink. Firms now reduce their prices when increasing output. Hence, the more elastic the demand, the less true profits will change in response to real incentives. The term multiplying the output elasticity is smaller than when we assume firms have no pricing power, and so we conclude that the presence of pricing power only strengthens our interpretation of observed bunching and makes our estimate of the evasion response based on the decomposition in (9) a lower bound.

IV. Context and Data

A. *Corporate Taxation: Minimum Tax Scheme*

The corporation tax is an important source of revenue in Pakistan and currently raises 2.5 percent of GDP, which constitutes about 25 percent of all federal tax revenues (World Bank 2009). The tax is remitted by about 20,000 corporations filing tax returns each year. The scale of noncompliance is suspected to be large in Pakistan, but credible evidence on the amount of corporate tax evasion has been lacking because of problems with data and methodology. The Federal Board of Revenue (FBR; 2013) reports an estimate of the corporate evasion rate equal to 45 percent but does not provide information on the estimation. A study by the World Bank (2009) estimates the evasion rate to be as high as 218 percent of actual corporate income tax payments, drawing on an input-output model for a selected group of sectors. It is this concern about corporate noncompliance that motivated policy makers in Pakistan to devise a tax scheme that ensures that every operational corporation pays a minimum amount of tax every year.[15] The minimum tax scheme, which has been in place since 1991, combines a tax rate τ_π on annual corporate profits (turnover minus deductible costs) with a smaller tax rate τ_y on annual corporate turnover, requiring each firm to assess both tax liabilities and pay, whichever is higher.[16]

As explained above, the minimum tax scheme implies that a firm's tax base depends on whether its profit rate (corporate profits as a share of turnover) is above or below a threshold equal to the tax rate ratio τ_y/τ_π. This profit rate threshold represents a kink point where the tax base and tax rate change discretely. The kink point varies across different groups of firms and across time. First, Pakistan offers a reduced profit tax rate for recently incorporated firms. All companies that register after June 2005, have no more than 250 employees, have annual turnover below Rs. 250 million, and have paid-up capital below Rs. 25 million are eligible for a lower profit tax rate. Second, both the profit tax rate and the turnover tax rate undergo changes during the time period we study. Table 1 (panel A) catalogs these variations across firms and over time, which we exploit in our empirical analysis. Importantly, the definitions of the tax bases to which these rates are applied remain the same for the entire period under consideration.

Finally, as shown in table 1, notice that the turnover tax is a significant feature of Pakistan's tax system: in the years we study, 50–60 percent of firms are liable for turnover taxation, and it accounts for as much as 60–75 percent of all corporate tax revenue.

B. Data

Our study uses administrative data from the FBR, covering the universe of corporate income tax returns for the years 2006–10.[17] Since July 2007,

[15] For example, in a testimony before the Federal Tax Ombudsman, Ikram ul Haq, who is a leading tax expert in Pakistan, has said that "the rationale for the levy of the alternate minimum tax was clear. So many inflated expenses are booked by taxpayers when filing returns that the tax base is drastically eroded and tax yields plummet to an intolerably low level. The only way out of this predicament is to resort to measures like enactment of the alternate minimum tax" (Federal Tax Ombudsman 2013).

[16] When the turnover tax is binding, firms are allowed to carry forward the tax paid in excess of the profit tax liability and adjust it against next year's profit tax liability, provided that the resulting net liability does not fall below the turnover tax liability for that year. Such adjustment, if not exhausted, can be carried forward for a period of up to 5 years (3 years in 2008 and 2009). In the data, we observe that only 1.3 percent of firms claim such carryforward, indicating either that firms are unaware of this option or that their profit tax liability net of carryforward drops below output tax liability, in which case carryforward cannot be claimed. In any case, the potential for carryforward attenuates the size of the minimum tax kink and works against the (strong) bunching that we find. We also exclude banks and financial firms, which face a standard tax rate of 38 percent in 2006, and 154 firms in sectors that were selectively given a lower turnover tax rate in 2010.

[17] In Pakistan, tax year t runs from July 1 of year t to June 30 of year $t + 1$.

TABLE 1
Descriptive Statistics

	2006	2007	2008	2009	2010
	A. Tax Variables				
Profit tax rate (high)	.35	.35	.35	.35	.35
Profit tax rate (low)	.2	.2	.2	.2	.25
Turnover tax rate	.005	.005	NA	.005	.01
Share of firms in turnover tax regime	.498	.561	NA	.532	.621
Share of revenue from turnover tax	.655	.713	NA	.581	.751
	B. Firm Characteristics (Means)				
Profits (million PKR)	1.1	.1	−.4	.3	1.1
	(.2)	(.2)	(.1)	(.1)	(.2)
Turnover (million PKR)	290.9	369.3	246.1	258.1	275.6
	(24.2)	(83.9)	(57.5)	(24.3)	(23.7)
Profit rate (profits/turnover)	2.08	2.06	1.95	2.2	2.46
	(.1)	(.06)	(.04)	(.05)	(.05)
Salary/turnover	.182	.200	.210	.217	.235
	(.005)	(.005)	(.004)	(.004)	(.005)
Interests/turnover	.019	.018	.018	.016	.018
	(.001)	(.002)	(.002)	(.001)	(.001)
Observations	8,604	14,587	20,485	19,944	19,909

Note.—The table presents descriptive statistics, focusing on tax variables (panel A) and firm characteristics (panel B). Rows 1–3 of panel A are based on Pakistan's corporate tax schedule. The remaining rows are based on administrative tax return data from the universe of tax-registered firms in Pakistan. All statistics are based on the raw data, excluding each year the top and bottom 5 percent tails in terms of profits. Standard errors are shown in parentheses. The low tax rate applies to firms that registered after June 2005, have no more than 250 employees, have annual sales of not more than Rs. 250 million, and have paid-up capital of not more than Rs. 25 million (all monetary figures are given in Pakistani rupees [PKR]).

electronic filing has been mandatory for all companies, and over 90 percent of the returns used in our study were filed electronically. Electronic filing ensures that the data have much less measurement error than is typically the case for developing countries. As far as we know, this is the first study to exploit corporate tax return data for a developing country. The filed returns are automatically subject to a basic validation check that uncovers any internal inconsistencies like reconciling tax liability with reported profit. Besides this validation check, the tax returns are considered final unless selected for audit.

Two aspects of the data are worth keeping in mind. First, our data set contains almost all active corporations. As corporations also act as withholding agents, deducting tax at source on their sales and purchases, it is almost impossible for an operational corporation not to file a tax return. FBR takes the view that registered corporations that do not file tax returns are nonoperational. Second, besides the corporations in our data, the population of firms in Pakistan includes both unincorporated firms subject to personal income taxation and informal firms operating outside the tax net. In general, corporate and personal income taxes may lead to shifting between the corporate and noncorporate sectors as well as between the formal and informal sectors (e.g., Waseem 2013), but we do not study these interesting effects here.[18]

Table 1 shows descriptive statistics for the full data set. We limit our empirical analysis to firms that report both profit and turnover and either the incorporation date or the profit tax liability, which are required for allocating firms to the high- and low-profit tax rate groups.[19] We also subject the data to a number of checks for internal consistency detailed in Appendix tables A2 and A3. Our final estimation data set contains 23,147 firm-year observations.

V. Empirical Results

This section presents the results of our empirical analysis, examining how firms respond to the minimum tax policy. We first present evidence that there is sharp bunching at the minimum tax kink as predicted by our analysis in Section III and that it is caused by the presence of the kink. We then use the observed bunching to estimate the magnitude of evasion responses.

A. *Bunching at Minimum Tax Kinks*

As shown in Section III, the type of minimum tax scheme observed in Pakistan should lead to excess bunching by firms around a threshold profit rate (profits as a share of turnover) equal to the ratio of the two tax rates, τ_y/τ_π. Figure 2 shows evidence that firms do indeed bunch around this minimum tax kink. The figure shows bunching evidence for different groups of firms (panels *a* and *b*) and different years (panels *c* and *d*), exploiting the variation in the kink across these samples. We plot the empirical density of the reported profit rate (profits as a percentage of turnover) in bins of approximately 0.2 percentage points. Panel *a* shows the density for high-rate firms (facing a profit tax rate of 35 percent) in the years 2006, 2007, and 2009 pooled together, since for those firms and years the minimum tax kink is at a profit rate threshold of $\tau_y/\tau_\pi = 0.5\%/35\% = 1.43\%$ (demarcated by a solid vertical line in the figure). The density exhibits large and sharp bunching around the kink point. Since there is no reason for firms to cluster around a profit rate of 1.43 percent other than the presence of the minimum tax scheme, this represents compelling evidence of a behavioral response to the scheme.

[18] Our bunching estimates capture intensive-margin responses conditional on being a corporate tax filer and are therefore not affected by incorporation or informality responses.

[19] Table A1 in the Appendix compares the firms we lose to those we are able to use on the basis of firm characteristics reported in the tax returns.

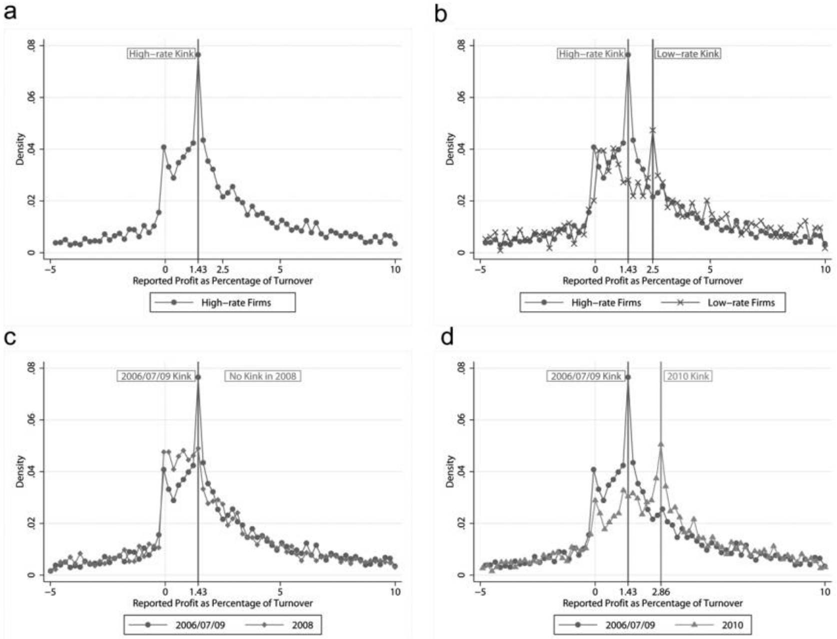

FIG. 2.—Bunching evidence: *a*, high-rate firms, 2006/7/9; *b*, low-rate versus high-rate firms, 2006/7/9; *c*, high-rate firms, 2008 versus 2006/7/9; *d*, high-rate firms, 2010 versus 2006/7/9. The figure shows the empirical density of the profit rate (reported profit as a percentage of turnover) for different groups of firms and time periods. Firms calculate their profit tax liability (based on tax rate τ_π) and their turnover tax liablity (based on tax rate τ_y) and pay whichever liability is larger. This minimum tax scheme creates a kink at a profit rate of τ_y/τ_π: firms are subject to profit taxation above the kink and turnover taxation below the kink. For high-rate firms in 2006/7/9 (panel *a*), $\tau_\pi = 0.35$ and $\tau_y = 0.005$, placing the kink at a profit rate of 1.43 percent. For low-rate firms in 2006/7/9 (panel *b*), $\tau_\pi = 0.20$ and $\tau_y = 0.005$, placing the kink at a profit rate of 2.5 percent. For high-rate firms in 2008 (panel *c*), the minimum tax scheme is abolished and all firms' profits are taxed at rate $\tau_\pi = 0.35$, so that there is no kink. For high-rate firms in 2010 (panel *d*), $\tau_\pi = 0.35$ and $\tau_y = 0.01$, placing the kink at a profit rate of 2.86 percent. Kink points are marked by vertical solid lines. The bin width is 0.214 in panels *a*–*c* and 0.204 in panel *d*, chosen to ensure that kink points are located at bin centers in all panels. The zero profit point is marked by a vertical dotted line. Color version available as an online enhancement.

Notice also that there is a modest amount of excess mass around the zero-profit point as many firms generate very little income (see, e.g., Burgstahler and Dichev [1997] for a discussion of excess mass at zero in profit distributions).[20]

Panels *b*–*d* provide identification checks ensuring that excess bunching at the minimum tax kink is indeed a response to the tax system (as opposed to a spurious property of the profit rate distribution) by ex-

[20] For comparison, fig. 1 in the online appendix reproduces panel *a* of fig. 2 in the text using the raw data before the consistency checks.

ploiting variation in the minimum tax kink across firms and over time. Panel *b* compares high-rate firms to low-rate firms during the years 2006, 2007, and 2009, when the latter group of firms face a reduced profit tax rate of 20 percent and therefore a minimum tax kink located at $\tau_y/\tau_\pi = 0.5\%/20\% = 2.5\%$. Besides the different location of the kink for low-rate firms, our model implies that the kink changes evasion incentives by less for low-rate firms (as the change in the evasion incentive $\tau\mu$ equals the profit tax rate τ_π) while it changes real incentives by the same amount (as the change in the real incentive $1 - \tau_E$ equals the turnover tax rate τ_y). Hence, we expect to see both that low-rate firms bunch in a different place and that the amount of bunching is smaller (if evasion is important), and this is precisely what panel *b* shows. Even though bunching is smaller for low-rate firms, it is still very clear and sharp. Outside the bunching areas around the two kinks, the low-rate and high-rate distributions are very close and exhibit the same (small) excess mass around zero.[21]

Panels *c* and *d* of figure 2 exploit time variation in the kink, focusing on the sample of high-rate firms. Panel *c* shows that excess bunching at 1.43 percent completely disappears in 2008 when the minimum tax regime (i.e., turnover tax below a profit rate of 1.43 percent) is removed. The 2008 density instead exhibits a larger mass of firms with profit rates between 0 and 1.43 percent. The distributions in panel *c* are consistent with our theoretical prediction that the introduction of an output tax below a profit rate threshold creates bunching coming from below. Finally, panel *d* shows that the bunch moves from 1.43 percent to 2.86 percent in 2010, when the doubling of the output tax rate shifts the kink. This change is accompanied by an overall decrease in the mass of firms with profit rates between 0 and 2, again illustrating that bunchers move to the kink from below.[22]

[21] The low-rate distribution is more noisy than the high-rate distribution because the former represents a much smaller fraction (about 22.9 percent) of the population of firms.

[22] The increase in the output tax rate τ_y in 2010 would be useful for identification if it were exogenous. As we described in Sec. III.A, exogenous variation in τ_y changes the real incentive without affecting the evasion incentive, allowing us to separately identify real output and evasion responses by comparing bunching at the minimum tax kink under a high-output tax (2010) and a low-output tax (2006/7/9). However, as shown in panel *d*, bunching in 2010 is smaller than in previous years, and so this method would yield a negative real output elasticity. The likely explanation is that the 2010 tax rate change cannot be viewed as exogenous. There are three possible reasons for this. First, there may be other confounding time changes that make bunching smaller in 2010, including optimization frictions that make it difficult for firms to respond quickly to a change in the location of the kink. Second, the increase in τ_y moves the kink point to a higher profit rate, and so bunchers are coming from a different part of the underlying profit rate distribution, which may feature lower evasion rates. Third, our baseline model in Sec. III.A did not allow for output evasion, an extension we considered in Sec. III.B. There we argued that the introduction of output evasion does not matter for any of our key results, but the one place where it would matter is in the implementation of the empirical strategy discussed here.

Taken together, the panels of figure 2 provide compelling evidence that firms respond to the incentives created by the minimum tax scheme. The substantial amount of bunching observed around kink points (which are associated with weak real incentives as explained above) suggests that evasion responses are quantitatively important.

B. Estimating Evasion Responses Using Bunching

This section presents estimates of excess bunching and uses our model to translate them into estimates of evasion responses. Following Chetty et al. (2011), we estimate a counterfactual density—what the distribution would have looked like without the kink—by fitting a flexible polynomial to the observed density, excluding observations in a range around the kink that is (visibly) affected by bunching. Denoting by d_j the fraction of the data in profit rate bin j and by π_j the (midpoint) profit rate in bin j, the counterfactual density is obtained from a regression of the following form:

$$d_j = \sum_{i=0}^{q} \beta_i (\pi_j)^i + \sum_{i=\pi_L}^{\pi_U} \gamma_i \cdot \mathbf{1}[\pi_j = i] + \nu_j, \tag{14}$$

where q is the order of the polynomial and $[\pi_L, \pi_U]$ is the excluded range. The counterfactual density is estimated as the predicted values from (14) omitting the contribution of the dummies in the excluded range, that is, $\hat{d}_j = \sum_{i=0}^{q} \hat{\beta}_i (\pi_j)^i$, and excess bunching is then estimated as the area between the observed and counterfactual densities in the excluded range, $\hat{B} = \sum_{j=\pi_L}^{\pi_U} (d_j - \hat{d}_j)$. Standard errors are bootstrapped by random resampling from the estimated residuals in (14).

Figure 3 compares the empirical density distributions to estimated counterfactual distributions (smooth solid lines) for the four samples examined in figure 2: high-rate firms in 2006/7/9 in panel *a*, low-rate firms in 2006/7/9 in panel *b*, high-rate firms in 2008 (placebo) in panel *c*, and high-rate firms in 2010 in panel *d*. In each panel, the solid vertical line represents the kink point while the dashed vertical lines demarcate the excluded range around the kink used in the estimation of the counterfactual.[23] To better evaluate the estimated counterfactuals, each panel also shows the empirical distribution for a comparison sample in light gray

This strategy relies on changes in the real incentive, but not the evasion incentive (which is true of variation in τ_y in the baseline model, but not in an extended model with output evasion). For all of these reasons, the 2010 change in τ_y does not help us separately estimate the real response elasticity.

[23] The excluded range $[\pi_L, \pi_U]$ is set to match the area around the kink in which the empirical density diverges from its smooth trend: four bins on either side of the kink in panels *a* and *c* and two bins on either side in panels *b* and *d*. The order of the polynomial q is five (seven for 2008), chosen so as to optimize the fit. Table A4 shows that the estimates are fairly sensitive to the choice of excluded range and polynomial degree in panel *a*, but less so in the other panels.

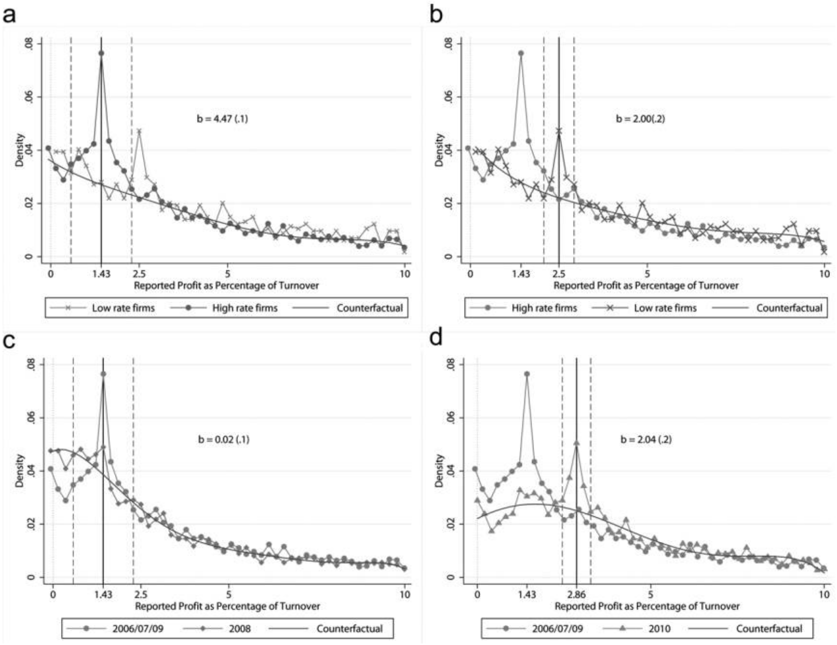

FIG. 3.—Bunching estimation: *a*, high-rate firms, 2006/7/9; *b*, low-rate firms, 2006/7/9; *c*, high-rate firms, 2008; *d*, high-rate firms, 2010. The figure shows the empirical density distribution of the profit rate (reported profit as a percentage of turnover, dotted dark graph), an empirical counterfactual density (dotted light graph), and the estimated counterfactual density (solid graph), for the different groups of firms and time periods considered in figure 2. The tax rate schedules and kink locations are explained in the notes to figure 2. The empirical counterfactual is the high-rate firms' density in 2006/7/9 for panels *b*–*d* and the low-rate firms' density in 2006/7/9 for panel *a*. The counterfactual density is estimated from the empirical density, by fitting a fifth-order polynomial (seventh-order for 2008), excluding data around the kink, as specified in equation (14). The excluded range is chosen as the area around the kink that is visibly affected by bunching. Kink points are marked by vertical solid lines; lower and upper bounds of excluded ranges are marked by vertical dashed lines. The zero profit point is marked by a dotted line. The bin size for the empirical densities is 0.214 (0.204 for 2010), so that the kink points are bin centers. Bunching *b* is the excess mass in the excluded range around the kink, in proportion to the average counterfactual density in the excluded range. Bootstrapped standard errors are shown in parentheses. Color version available as an online enhancement.

(low-rate firms in panel *a*, high-rate firms in panel *b*, and 2006/7/9 in panels *c* and *d*). The observation that in all cases the empirical distribution for our comparison sample lines up well with the estimated counterfactual, particularly around the kink, provides a further validation of our estimates.

The figure also displays estimates of excess bunching scaled by the average counterfactual density around the kink, that is, $b = \hat{B}/E(\hat{d}_j \mid j \in [\pi_L, \pi_U])$. In general, these bunching estimates are large and strongly statistically significant, except in the placebo analysis of panel *c*, where bunching is close to zero and insignificant. Excess bunching is larger for high-rate firms in 2006/7/9 ($b = 4.47$ (0.1)) than for low-rate firms in the same period ($b = 2.00$ (0.2)), consistent with the fact that a lower profit tax rate implies a smaller change in the evasion incentive at the kink. Furthermore, excess bunching by high-rate firms is larger during the years 2006/7/9 than in year 2010 ($b = 2.04$ (0.1)), possibly because optimization frictions prevent some firms from responding to the change in the location of the kink in the short run.

Table 2 converts our bunching estimates into evasion responses using the methodology developed in Section III. As shown earlier, the amount of bunching translates to a reported profit rate response via the relationship $\Delta\hat{\pi} = B/f_0(\tau_y/\tau_\pi) \simeq b \times$ bin width,[24] and this profit rate response is in turn linked to the combination of real output and evasion responses via equation (9):

$$\Delta\hat{\pi} \simeq \frac{\tau_y^2}{\tau_\pi}\varepsilon_y - \frac{d(\hat{c} - c)}{y}.$$

The table shows estimates of excess bunching b in column 1, the profit rate response $\Delta\hat{\pi}$ in column 2, the real output elasticity ε_y assuming zero evasion in column 3, and the evasion response assuming different real output elasticities $\varepsilon_y \in \{0, 0.5, 1, 5\}$ in columns 4–7. Evasion responses are reported as percentages of taxable profits (evasion rate responses) instead of percentages of output in equation (9). Evasion rates in terms of output are easily converted to evasion rates in terms of profits, using the fact that $(y - \hat{c})/y = \tau_y/\tau_\pi$ at the kink. The different rows of the table show results for the main subsamples considered in the bunching figures.

The following main findings emerge from the table. First, in a model without evasion, the bunching we observe implies phenomenally large real output elasticities, ranging from 15 to 134 across the different samples. These elasticities are all far above the upper bound of the range of values that can be considered realistic, and so we can comfortably reject that model.[25] The reason for the large elasticities in this model is the combination of large observed bunching and the tiny variation in real incentives at the kink. Second, when we allow for tax evasion in the model,

[24] Since b equals bunching divided by the counterfactual density in discrete bins, we have to multiply b by bin width to obtain the profit rate response. The bin width underlying b is 0.214 percentage points for most estimates, and so bin width = 0.00214.

[25] For example, Gruber and Rauh (2007) estimate that the elasticity of corporate taxable income with respect to the effective marginal tax rate in the United States is 0.2. Taking that estimate at face value in the Pakistani context, with all the caveats that entails, would imply that

$$0.2 = \frac{d\text{CTI}}{d\tau_E}\frac{\tau_E}{d\text{CTI}} = \frac{\partial\text{CTI}/\partial y}{\text{CTI}/y}\varepsilon_y,$$

and so even a real output elasticity of 15 would require marginal taxable profits to be 1.33 percent of average taxable profits to be reconcilable with their estimate.

TABLE 2
Estimating Evasion Responses

	Observed Responses		Model Without Evasion	Model With Evasion			
				Evasion Rate Response			
	Bunching (b) (1)	Profit Rate Response ($\Delta\hat{\pi}$) (2)	Output Elasticity (ε_y) (3)	$\varepsilon_y = 0$ (4)	$\varepsilon_y = .5$ (5)	$\varepsilon_y = 1$ (6)	$\varepsilon_y = 5$ (7)
High-rate firms, 2006/7/9	4.47	1.0	134.2	67.1	66.9	66.6	64.6
	(.1)	(.03)	(4.0)	(2.0)	(2.0)	(2.0)	(2.0)
Low-rate firms, 2006/7/9	2.00	.4	34.3	17.1	16.9	16.6	14.6
	(.2)	(.04)	(3.1)	(1.6)	(1.6)	(1.6)	(1.6)
High-rate firms, 2010	2.04	.4	14.6	14.6	14.1	13.6	9.6
	(.1)	(.03)	(1.1)	(1.1)	(1.1)	(1.1)	(1.1)

Note.—This table presents bunching, elasticity, and evasion estimates for the subsamples considered in panels *a*, *b*, and *d* of fig. 3. Column 1 reproduces the bunching estimate *b*, based on estimating eq. (14). Bunching *b* is the excess mass in the excluded range around the kink, in proportion to the average counterfactual density in the excluded range. Column 2 presents an estimate of the profit rate response (in percent) associated with *b*, based on the relationship $\Delta\hat{\pi} = B/f_0(\tau_y/\tau_\pi) \simeq b \times$ bin width. Column 3 presents estimates of the real output elasticity ε_y for the model without evasion. This model is based on the assumption that bunching is purely due to a real output response. The elasticity ε_y is estimated using the relationship $\Delta\pi = [c/y - c'(y)]dy/y \simeq (\tau_y^2/\tau_\pi)\varepsilon_y$. Columns 4–7 present estimates of the evasion response as a percentage of taxable profits (evasion rate responses) for the model with evasion. This model allows for bunching to be driven by both evasion and real output response. The evasion response estimates are based on $\Delta\hat{\pi} = [\hat{c}/y - c'(y)]dy/y - [d(\hat{c} - c)/y] \simeq (\tau_y^2/\tau_\pi)\varepsilon_y - (\tau_y/\tau_\pi)[d(\hat{c} - c)/(y - \hat{c})]$, assuming different real output elasticities $\varepsilon_y \in \{0, 0.5, 1, 5\}$. Bootstrapped standard errors are shown in parentheses.

it becomes possible to reconcile observed bunching with reasonable values of the real output elasticity combined with large (but not implausible) evasion responses. Column 3 provides an upper bound on the evasion response, assuming a zero real output response. In this case, the evasion response ranges from 14.7 percent to 67.1 percent of profits across the different populations, with high-rate firms in 2006/7/9 featuring the largest response. Third, the evasion estimates are very robust to the real output elasticity even though we allow for elasticities up to 5, much higher than the empirical literature suggests is justified. The reason for this robustness is again that real incentives at the kink are extremely small. Hence, while we cannot separately identify both real and evasion responses using the minimum tax kink, we can provide very tight bounds on the evasion response due to the particular set of incentives provided by the minimum tax kink.

The evasion estimates in table 2 use our baseline model with only cost evasion, but the results are very robust to allowing for output evasion as in the conceptual model of Section III.B. In the model with both cost and output evasion, bunching identifies approximately the aggregate evasion reduction when switching from profit to turnover taxation. To span the range of possible estimates, table 1 in the online appendix considers the opposite extreme in which only output can be evaded (using eq. [11] assuming zero cost evasion) and shows that the evasion estimates are virtually unaffected.

C. *Tax Evasion versus Lazy Reporting*

In Section III.B, we considered a model with filing costs in which bunching at the minimum tax kink is driven not only by *cost overreporting* under the profit tax but also by *cost underreporting* under the turnover tax. As equation (12) shows, in the presence of filing costs, using equation (9) to infer evasion responses from bunching will conflate evasion responses with filing responses and overestimate evasion. This section presents the results of an identification check that suggests that lazy reporting is not a key confounder in practice.[26]

[26] It is worth noting that underreporting costs within the turnover tax regime is in fact a form of noncompliance under Pakistan's filing rules. The relevant tax law (Income Tax Ordinance of 2001) states in sec. 114 that every company must file a return that "fully states all the relevant particulars or information as specified in the form of return, including a declaration of the records kept by the taxpayer." The law then states in secs. 120 and 182 what will happen if the return is incomplete, including the penalties associated with such noncompliance. While these legal provisions obviously do not rule out noncompliance in the form of cost underreporting, they do raise the bar for the level of filing costs necessary to induce such cost underreporting.

Our identification test exploits the extremely detailed nature of our data, which include every single line item on the corporate tax return. The return includes a total of 42 cost line items, ranging from very common items such as "salaries" and "stationery" to rare items such as "Zakat" (an Islamic charitable contribution), "stocks/stores/spares/fixed assets written off as obsolete," and "imported finished goods." Since we observe these 42 cost items and all the firms in our data report some but not all of the items, we can use the number of items reported as a measure of q in the model in Section III.B and test whether q responds to the minimum tax kink.

The test is presented in figure 4, which shows four panels for high-rate versus low-rate firms (top vs. bottom panels) and for all cost items versus rare cost items (left vs. right panels). Each panel shows the fraction of cost items filed on the y-axis and the profit rate (with the minimum tax kink demarcated by a vertical line) on the x-axis. The dots show tax years 2006/7/9 (when the kink was in place), while the crosses represent tax year 2008 (when the kink was temporarily abolished). The right panels consider "rare cost items" defined as items that are below the median in the distribution of filing probabilities in the full population of firms. We consider rare cost items separately since our model in Section III.B suggests that the turnover tax is more likely to affect items that are seldom filed in the first place (items with high filing costs $f(r)$).

The following findings emerge from the figure. First, the dotted series shows no sign of a discontinuity at the kink (in particular, it is not discretely lower on the left) in any of the panels. Second, in panels a and b the dotted series is weakly decreasing with profit rates; that is, firms in the turnover tax regime tend to report more cost items. Third, to control for unobserved heterogeneity in filing behavior across firms (which may be correlated with being in the turnover tax regime but not causally driven by the turnover tax), we add the 2008 series when there was no minimum tax kink. In all four panels, the two series are virtually indistinguishable both above and below the kink. Fourth, each panel reports difference-in-differences (DD) estimates comparing treated firms (below the kink) to control firms (above the kink) over time. The regression specification is shown in the note to the figure. We show two estimates: DD is based on the full population of firms, while DD^{near} is based on a subsample of firms near the kink (those with a reported profit rate within 1.5 percentage points of the kink). Both DD estimates are very close to zero and statistically insignificant in every panel.

Overall, the evidence presented in figure 4 shows no effect on the number of line items reported, which is inconsistent with the lazy reporting story. This lends support to our interpretation that bunching at the minimum tax kink is driven by deliberate evasion (greater amount of cost per

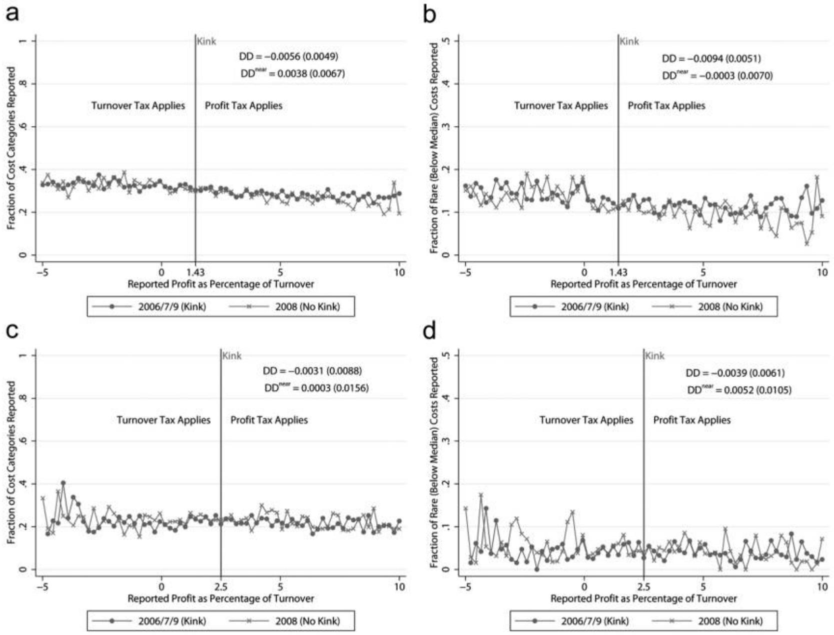

FIG. 4.—Usage of cost categories varies smoothly across the kink: *a*, high-rate firms' usage of all cost categories; *b*, high-rate firms' usage of rare cost categories; *c*, low-rate firms' usage of all cost categories; *d*, low-rate firms' usage of rare cost categories. The figures show how firms' usage of the 42 cost categories available on the tax return varies with their reported profit rate. The dots show the fraction of costs used by firms in 2006, 2007, and 2009: years in which the turnover tax is in place below the kink. The crosses show the fraction of costs used in 2008 when there is no turnover tax. Panels *a* and *c* show the fraction of the 42 cost categories used by high-rate and low-rate firms, respectively. Panels *b* and *d* show the fraction of the 21 least-used categories used by the high-rate and low-rate firms, respectively. The figures also show the coefficient from the following difference-in-differences regression: $f_{it} = \alpha + \beta I\{\text{Turnover Tax}\}_i + \gamma I\{2006/7/9\}_t + \text{DD}I\{\text{Turnover Tax}\}_i \times I\{2006/7/9\}_t + \varepsilon_{it}$, where f_{it} is the fraction of costs reported by a firm i in year t; $I\{\text{Turnover Tax}\}_i$ is a dummy for reporting a profit rate below the kink (1.43 percent for the high-rate firms and 2.5 percent for the low-rate firms); $I\{2006/7/9\}_t$ is a dummy for observations from the years 2006, 2007, and 2009; and ε_{it} is an error term. The coefficient DD in the figures is obtained from a regression using all available firms, while the coefficient DD^{near} estimates the regression on the subsample of firms with reported profit rates within 1.5 percentage points of the kink. Color version available as an online enhancement.

line item under profit taxation) as opposed to systematic reporting omissions (fewer line items under turnover taxation).

VI. Numerical Analysis of Tax Policy Implications

This section links our empirical results to the stylized model in Section II. We first provide some reduced-form conclusions based on our optimal tax rule (5) and then put more structure on firms' production and evasion decisions to assess the welfare implications of different tax regimes. The analysis will be based on our partial equilibrium model in Section II.A. Our empirical estimates do not capture general equilibrium effects, and these effects would be particularly hard to calibrate; this would require us to measure production chains in a multisector model, and our administrative data contain no information on this. While our two-sector model in Section II.B shows that cascading and incidence effects in general equilibrium have offsetting effects on optimal policy, here we quantify how far the partial equilibrium channel by itself can move the optimal policy away from production efficiency.

A. *Welfare Analysis for Uniform Tax Regimes*

This section considers a uniform tax rate and tax base on all firms. Without imposing additional structure on firms' output and evasion responses, we can evaluate the desirability of local changes in the tax policy based on the optimal tax rule in proposition 1. In particular, one can increase welfare by broadening the base and lowering the rate as long as

$$\frac{\tau}{1-\tau}\cdot\frac{\partial\tau_E}{\partial\tau}(\mu) < G(\mu)\cdot\frac{\varepsilon_{\hat{c}-c}}{\varepsilon_y},$$

with both sides evaluated for the policy in place. The left-hand side of the inequality reflects the effective distortion of real production and is fully determined by our theoretical model. The right-hand side reflects the importance of evasion responses relative to production responses. We can rewrite this in terms of the estimated evasion response at the minimum tax kink,

$$G(\mu)\cdot\frac{\varepsilon_{\hat{c}-c}}{\varepsilon_y} \simeq -\frac{d(\hat{c}-c)}{\hat{\Pi}}\Big/\varepsilon_y,$$

where we have used $d(\tau\mu)/(\tau\mu) = -1$ at the minimum tax kink. This is an approximation as the change in evasion incentives at the kink is not marginal. This approximation is exact if the evasion cost function is isoelastic, which we assume in our numerical calibration below.

For the firms facing a profit tax rate of $\tau = .35$, our empirical analysis provides an estimate of the right-hand side equal to 1.34 when assuming $\varepsilon_y = .5$ (i.e., when using the evasion rate response of 66.9 percent shown in table 2). This is always larger than the left-hand side, which ranges from 0 under a pure profit tax base to $\tau/(1-\tau) = .35/.65 = .54$ under a pure turnover tax base. Tax evasion by high-rate firms in response to the current profit tax is thus too high relative to the effective tax wedge, indicating that welfare could be increased by broadening the tax base and decreasing the tax rate. For the firms facing a lower profit tax rate of $\tau = .20$, we find a lower estimate for the right-hand side of .34. While the tax evasion response is substantially smaller for these firms, it still exceeds the upper bound for the left-hand side $(\tau/[1 - \tau]) = .20/.80 = .25)$, justifying a move away from the profit tax base even at the lower 20 percent rate. To fully determine the optimal tax base and rate, the optimal tax rule needs to be considered jointly with a revenue requirement or a constraint on firms' profits.

By putting more structure on the firms' production and evasion technologies, we can evaluate the welfare gains from switching between a pure profit tax and a pure turnover tax, as well as the gains from an optimal tax system that generally lies in between the two extremes. We consider firms operating with iso-elastic production and evasion cost functions. The parameters of firm-specific production functions are calibrated to replicate the empirical distributions of turnover and costs, while the parameters of the evasion cost function are calibrated to match our evasion response estimates for high-rate and low-rate firms. Our calibration is based on year 2008, during which the minimum tax regime was not in place and all firms were subject to a profit tax. Full details are provided in Section C in the Appendix.

The results of the calibration are shown in table 3, with panel A considering a pure turnover tax and panel B considering the optimal tax system. Assuming a real output elasticity of 0.5, we find that corporate tax revenues increase by 74 percent when switching from a pure profit tax to a pure turnover tax and adjusting the tax rate so that aggregate (after-tax) profits remain the same. Given a profit tax rate of 35 percent, this requires a turnover tax rate of 0.5 percent, which coincides with the actual turnover tax rate in Pakistan. Note that revenue increases, holding aggregate profits constant, represent welfare gains. By comparing panels A and B, we see that a pure turnover tax realizes virtually all of the welfare gain from setting the tax base and rate at their optimal levels. At an output elasticity of 0.5, the optimal policy is characterized by $\tau = .009$ and $\mu = .522$ as this maximizes corporate tax revenue while keeping aggregate profits constant compared to the benchmark profit tax.[27] As shown in the table, the dominance of a pure turnover tax over a pure profit tax is robust for a very wide range of output elasticities. For example, the revenue gain from switching bases is still 66 percent for an arguably extreme output elasticity of 10. The optimal tax base moves closer to profits as the real output elasticity increases, but most of the po-

[27] While these calibrations use our baseline model with only cost evasion, the results are again robust to allowing for output evasion. For example, when recalibrating the model assuming that only output can be evaded, the revenue gain from switching to a pure turnover tax is 78 percent (instead of 74 percent), and again this realizes virtually all the gain of the optimal tax system characterized by $\tau = .006$ and $\mu = .23$.

TABLE 3
Numerical Tax Policy Analysis

	A. Pure Turnover Tax			B. Optimal Tax		
Output Elasticity (ε_y)	Revenue Gain (%) (1)	Tax Base (μ) (2)	Tax Rate (τ) (3)	Revenue Gain (%) (4)	Tax Base (μ) (5)	Tax Rate (τ) (6)
.5	74	0	.005	76	.522	.009
1	73	0	.005	76	.706	.015
5	70	0	.005	75	.889	.037
10	66	0	.005	75	.944	.067
30	62	0	.005	77	.986	.170

Note.—This table presents a numerical evaluation of the tax policy changes discussed in Sec. VI. For an assumed output elasticity, we calibrate firm-specific production parameters and cost evasion parameters to replicate the empirical distributions of turnover and costs and to match our evasion response estimates. All details of the calibration are in Sec. C of the Appendix. In panel A, we simulate a switch from a pure profit tax ($\tau = 0.35$; $\mu = 1$) to a pure turnover tax, adjusting the tax rate so that aggregate (after-tax) profits remain the same. In panel B, we simulate a switch from the same pure profit tax ($\tau = 0.35$; $\mu = 1$) to the optimal tax policy, maximizing tax revenues without reducing aggregate profits. Columns 1 and 4 show the implied tax revenue gain for the pure turnover tax and the optimal tax, respectively. Columns 2 and 5 show the corresponding tax base parameters. Columns 3 and 6 show the corresponding tax rates. The results illustrate that turnover taxation dominates profit taxation and realizes almost all potential welfare gains from setting the tax base and rate optimally.

tential welfare gain can still be realized by switching to the simpler turnover tax.

B. *Uniform versus Minimum Tax Regimes*

The preceding analysis considered a uniform tax base on all firms as opposed to a minimum tax scheme imposing turnover taxation on some firms and profit taxation on others. While it is standard to exploit quasi-experimental variation coming from nonuniform policies to estimate parameters that inform the optimal uniform policy,[28] the desirability of minimum tax schemes is an interesting question by itself given the ubiquity of such policies.

Conceptually, the reason why it may be optimal to allow the tax base μ to vary across firms is that the underlying tax base determinants characterized in equation (5) vary by firm. In particular, if either the evasion elasticity $\varepsilon_{\hat{c}-c}$ is larger or the real output elasticity ε_y is smaller among firms with lower reported profit rates, then it is optimal to set a lower μ among such firms. Interestingly, the empirical evidence in figure 3 and table 2 lends support to the idea that the evasion elasticity is declining in the reported profit rate: the evasion response by high-rate firms in

[28] For example, exploiting tax reforms or kink points in piecewise linear income tax systems to estimate sufficient statistics for optimal linear income taxes (see, e.g., Saez, Slemrod, and Giertz 2012).

2006/7/9 is much larger than the evasion response by low-rate firms in those years or by high-rate firms in 2010, and the minimum tax kink for the former group of firms is located at a much lower point in the profit rate distribution than for the other groups. We of course need to be cautious when interpreting such cross-sectional and time variation in bunching estimates, but the large differences suggest that evasion may be less of an issue for firms reporting larger profit rates.

While these considerations can justify an increasing μ in profit rates, the minimum tax scheme with μ jumping discretely from zero to one at a cutoff is an extreme policy that is unlikely to be optimal in an unconstrained policy setting. To justify such simple schemes, we would have to allow for costs of administration and complexity that make complicated, nonlinear μ schedules undesirable. These issues are of course extremely relevant in a setting with limited tax capacity.

VII. Conclusion

In this paper we have studied the trade-off between preserving production efficiency and preventing the corrosion of revenues due to tax evasion faced by governments with limited tax capacity. Our focus has been a production inefficient policy commonly observed in developing countries: taxing firms on the basis of turnover rather than profits. In contrast to models without evasion in which the optimal tax base is pure profits (preserving production efficiency), in the presence of evasion the optimal tax base sacrifices some production efficiency in order to curtail evasion levels. Our optimality conditions relate the choice of tax base to the elasticities of real production and tax evasion with respect to the tax wedge on each margin. A pure turnover tax may be better than a pure profit tax in terms of social welfare, although in general the social optimum lies in between the two extremes.

To study this empirically, we have developed a quasi-experimental approach based on a minimum tax scheme that is ubiquitous in the developing world: taxing each firm on either profits or turnover, depending on which tax liability is larger. We have shown that such schemes can be used to estimate tight bounds on the evasion response to switching from profits to turnover taxation. Using administrative tax records on corporations in Pakistan, we estimate that a switch from profit taxation to turnover taxation reduces evasion levels by up to 60–70 percent of corporate income. Linking these estimates back to our theoretical framework, we find that the optimal tax system has a base that is far broader than profits and that a switch to pure turnover taxation (although not the social optimum) may create a revenue gain of 74 percent without reducing aggregate profits. This welfare gain does not incorporate general equilibrium effects—including those coming from cascading effects of turnover taxation—which we show have offsetting impacts on the welfare analysis and depend on parameters that we do not estimate. Future work will hopefully provide evidence on those general equilibrium aspects in order to provide a stronger foundation for the policy advice given to developing countries, building on the approach we have developed here.

Appendix A

A. Proof of Proposition 1

Since firms are optimizing their choices of y and $\hat{c}$, the envelope theorem implies that we can write the changes in welfare from changes in τ, μ as

$$\frac{\partial W}{\partial \tau} = (\lambda - 1)\frac{\partial T(y,\ \hat{c})}{\partial \tau} + \lambda\left[\frac{\partial T(y,\ \hat{c})}{\partial y}\frac{\partial y}{\partial \tau} + \frac{\partial T(y,\ \hat{c})}{\partial \hat{c}}\frac{\partial \hat{c}}{\partial \tau}\right],$$

$$\frac{\partial W}{\partial \mu} = (\lambda - 1)\frac{\partial T(y,\ \hat{c})}{\partial \mu} + \lambda\left[\frac{\partial T(y,\ \hat{c})}{\partial y}\frac{\partial y}{\partial \mu} + \frac{\partial T(y,\ \hat{c})}{\partial \hat{c}}\frac{\partial \hat{c}}{\partial \mu}\right].$$

Rewriting $T(y,\ \hat{c})$ as $\tau(y - \mu[(\hat{c} - c) + c])$, note that $\partial T/\partial y = \tau[1 - \mu c'(y)] = \tau_E$ using the firm's optimality condition for y and that $\partial T/\partial(\hat{c} - c) = -\tau\mu$. Denote the (normalized) mechanical welfare effects of τ and μ by

$$M_\tau \equiv \frac{\partial T(y,\ \hat{c})}{\partial \tau} \times \frac{\lambda - 1}{\lambda} = (y - \mu\hat{c}) \times \frac{\lambda - 1}{\lambda} \geq 0$$

and

$$M_\mu \equiv \frac{\partial T(y,\ \hat{c})}{\partial \mu} \times \frac{\lambda - 1}{\lambda} = -\tau\hat{c} \times \frac{\lambda - 1}{\lambda} \leq 0,$$

so that the total welfare effects of changing τ, μ can be written as

$$\frac{\partial W}{\partial \tau}\bigg/\lambda = M_\tau + \tau_E\frac{\partial y}{\partial \tau} - \tau\mu\frac{\partial(\hat{c} - c)}{\partial \tau}, \tag{A1}$$

$$\frac{\partial W}{\partial \mu}\bigg/\lambda = M_\mu + \tau_E\frac{\partial y}{\partial \mu} - \tau\mu\frac{\partial(\hat{c} - c)}{\partial \mu}. \tag{A2}$$

For $\mu = 1$, an increase in the tax base has a second-order negative impact on production efficiency but a first-order positive impact on evasion reduction, that is,

$$\frac{\partial W}{\partial \mu}\bigg/\lambda = M_\mu - \tau\frac{\partial(\hat{c} - c)}{\partial \mu} < 0.$$

This result also holds for $\lambda = 1$, in which case $M_\mu = 0$.

For $\mu = 0$, a decrease in the tax base has a second-order negative impact on evasion reduction but a first-order positive impact on production efficiency. Notice that

$$\frac{\partial W}{\partial \mu}\Big/\lambda = M_\mu + \tau\frac{\partial y}{\partial \mu} > 0$$

if M_μ is sufficiently small. However, since the impact on evasion is of second order, we can use the same argument as before to argue that a tax-neutral increase in μ and τ, for a given y, will increase y and thus increase welfare, starting from $\mu = 0$.

To characterize the relation between the tax rate τ and the tax base μ, consider a joint increase $d\mu$ and $d\tau = [\tau\hat{c}/(y - \mu\hat{c})]d\mu$ such that the mechanical welfare effects M_τ and M_μ cancel out. The welfare effect of this change thus depends on the responses in y and $\hat{c} - c$. That is,

$$\begin{aligned} dW/\lambda &= \left(\frac{\partial W}{\partial \tau}\Big/\lambda\right)d\tau + \left(\frac{\partial W}{\partial \mu}\Big/\lambda\right)d\mu \\ &= \tau_E\frac{\partial y}{\partial \tau_E}\left(\frac{\partial \tau_E}{\partial \tau}d\tau + \frac{\partial \tau_E}{\partial \mu}d\mu\right) - \tau\mu\frac{\partial(\hat{c} - c)}{\partial \tau\mu}\left(\frac{\partial \tau\mu}{\partial \tau}d\tau + \frac{\partial \tau\mu}{\partial \mu}d\mu\right) \\ &= \tau_E\frac{\partial y}{\partial \tau_E}\frac{\partial \tau_E}{\partial \tau}\left(\frac{\tau\hat{c}}{y - \mu\hat{c}} + \frac{\partial \tau_E/\partial \mu}{\partial \tau_E/\partial \tau}\right)d\mu - \tau\mu\frac{\partial(\hat{c} - c)}{\partial \tau\mu}\left(\mu\frac{\tau\hat{c}}{y - \mu\hat{c}} + \tau\right)d\mu. \end{aligned}$$

Rewriting this in terms of elasticities, using

$$\frac{\partial \tau_E/\partial \mu}{\partial \tau_E/\partial \tau} = -\frac{\tau(1 - \tau)}{1 - \mu},$$

we find

$$dW/\lambda = \left[\frac{\tau}{1 - \tau}\frac{\partial \tau_E}{\partial \tau}\hat{\Pi}(y,\ \hat{c})\varepsilon_y - (\hat{c} - c)\varepsilon_{\hat{c}-c}\right]\frac{\tau y}{y - \mu\hat{c}}d\mu.$$

Notice that $dW/\lambda = 0$ is required for the initial level of τ and μ to be optimal, and so the expression in the proposition follows.

B. *Proof of Proposition 2*

In general equilibrium, welfare is given by $W = \Pi_A + \Pi_B + \lambda(T_A + T_B - R)$, and so the total welfare effect of changing τ (by an envelope argument similar to the proof of proposition 1) is

$$\begin{aligned} \frac{\partial W}{\partial \tau}\Big/\lambda &= \tilde{M}_\tau + \tau[p_A(1 - \mu) - \mu w]\frac{\partial y_A}{\partial \tau} + \tau\frac{\partial y_B}{\partial \tau} - \tau\mu w\frac{\partial l_B}{\partial \tau} - \tau\mu\frac{\partial(\hat{c} - c)}{\partial \tau} \\ &= \tilde{M}_\tau + \tau F'_{y_A}(1 - \tau_E)[(1 - \mu) - \mu(1 - \tau_E)]\frac{\partial y_A}{\partial \tau} \\ &\quad + \tau\frac{\partial y_B}{\partial \tau} - \tau\mu F'_{l_B}(1 - \tau_E)\frac{\partial l_B}{\partial \tau} - \tau\mu\frac{\partial(\hat{c} - c)}{\partial \tau} \\ &= \tilde{M}_\tau + \tau_E[1 + \alpha(\mu)]\frac{\partial y_B}{\partial \tau} - \tau\mu\frac{\partial(\hat{c} - c)}{\partial \tau}, \end{aligned} \tag{A3}$$

with

$$\alpha(\mu) = \mathrm{MRTS}_{l_B,y_A} \Big/ \left[1 + \mathrm{MRTS}_{l_B,y_A}\left(\frac{\partial l_B}{\partial \tau_E} \Big/ \frac{\partial y_A}{\partial \tau_E}\right)\right].$$

The second equality follows by inserting the final sector firm's optimality conditions and the third by noting that $\partial y_B/\partial\tau = F'_{l_B}\partial l_B/\partial\tau + F'_{y_A}\partial y_A/\partial\tau$ and that the final sector firm sets $\mathrm{MRTS}_{l_B,y_A} = 1-\tau_E$. The mechanical effect in the general equilibrium setting now also incorporates an incidence effect:

$$\begin{aligned}\tilde{M}_\tau\frac{\lambda}{\lambda-1} &= \underbrace{p_Ay_A + y_B - \mu\hat{c}}_{\text{declared profits taxed more}} + \underbrace{\tau y_A(1-\mu)\frac{\partial p_A}{\partial\tau}}_{\text{incidence effect}} \\ &= p_Ay_A + y_B - \mu\hat{c} + \frac{\tau}{\tau_E}p_Ay_A(1-\mu)\varepsilon_{p_A}\frac{\partial\tau_E}{\partial\tau},\end{aligned} \tag{A4}$$

where $\varepsilon_{p_A} \equiv (\partial p_A/\partial\tau_E)/(\tau_E/p_A)$ is the elasticity of the intermediate good's price with respect to the effective tax rate. Similarly, the welfare effect of changing μ is

$$\begin{aligned}\frac{\partial W}{\partial\mu}\Big/\lambda &= \tilde{M}_\mu + \tau[p_A(1-\mu) - \mu w]\frac{\partial y_A}{\partial\mu} + \tau\frac{\partial y_B}{\partial\mu} - \tau\mu w\frac{\partial l_B}{\partial\mu} - \tau\mu\frac{\partial(\hat{c}-c)}{\partial\mu} \\ &= \tilde{M}_\mu + \tau_E[1+\alpha(\mu)]\frac{\partial y_B}{\partial\mu} - \tau\mu\frac{\partial(\hat{c}-c)}{\partial\mu},\end{aligned} \tag{A5}$$

where the mechanical effect is

$$\begin{aligned}\tilde{M}_\mu\frac{\lambda}{\lambda-1} &= -\underbrace{\tau\hat{c}}_{\text{declared costs taxed less}} + \underbrace{y_A\tau(1-\mu)\frac{\partial p_A}{\partial\mu}}_{\text{incidence effect}} \\ &= -\tau\hat{c} + \frac{\tau}{\tau_E}p_Ay_A(1-\mu)\varepsilon_{p_A}\frac{\partial\tau_E}{\partial\mu}.\end{aligned} \tag{A6}$$

To characterize the relation between the tax rate τ and the tax base μ, consider a joint increase $d\mu$ and $d\tau$ such that the mechanical welfare effects $\tilde{M}_\tau$ and $\tilde{M}_\mu$ cancel out:

$$d\tau = \frac{\tau\hat{c} - \dfrac{\tau}{\tau_E}p_Ay_A(1-\mu)\varepsilon_{p_A}\dfrac{\partial\tau_E}{\partial\mu}}{p_Ay_A + y_B - \mu\hat{c} + \dfrac{\tau}{\tau_E}p_Ay_A(1-\mu)\varepsilon_{p_A}\dfrac{\partial\tau_E}{\partial\tau}}d\mu.$$

The welfare effect of this change is then

$$\begin{aligned}dW/\lambda &= \left(\frac{\partial W}{\partial\tau}\Big/\lambda\right)d\tau + \left(\frac{\partial W}{\partial\mu}\Big/\lambda\right)d\mu \\ &= -\tau_E[1+\alpha(\mu)]\frac{\partial y_B}{\partial(1-\tau_E)}\frac{\partial\tau_E}{\partial\tau}\left(\frac{d\tau}{d\mu} + \frac{\partial\tau_E/\partial\mu}{\partial\tau_E/\partial\tau}\right)d\mu \\ &\quad -\tau\mu\frac{\partial(\hat{c}-c)}{\partial\tau\mu}\left(\mu\frac{d\tau}{d\mu} + \tau\right)d\mu.\end{aligned}$$

Rewriting in terms of elasticities using

$$\frac{\partial\tau_E/\partial\mu}{\partial\tau_E/\partial\tau} = -\frac{\tau(1-\tau)}{1-\mu},$$

we find

$$\begin{aligned}\frac{dW}{\lambda} &= \left\{\frac{\tau}{1-\tau}[1+\alpha(\mu)]\frac{\partial\tau_E}{\partial\tau}\varepsilon_y\beta\hat{\Pi} - (\hat{c}-c)\varepsilon_{\hat{c}-c}[1+(1-\beta)\varepsilon_{p_A}]\right\}\\ &\quad\times\frac{\tau(y_B+p_Ay_A)}{p_Ay_A+y_B-\mu\hat{c}+\dfrac{\tau}{\tau_E}p_Ay_A(1-\mu)\varepsilon_{p_A}\dfrac{\partial\tau_E}{\partial\tau}}d\mu,\end{aligned}$$

and setting the term in braces to zero yields the expression in the proposition.

C. *Calibration Details*

This section provides further details on the modeling assumptions and the different steps of the calibration underlying the numerical welfare analysis in Section VI. We assume a production function with constant elasticity ε_y,

$$y_i(c) = A_i(c-F_i)^{\varepsilon_y/(1+\varepsilon_y)}\Big/\frac{\varepsilon_y}{1+\varepsilon_y},$$

where A_i is a firm-specific scale parameter and F_i captures the firm's fixed costs, and an evasion cost function with constant elasticity $\varepsilon_{\hat{c}-c}$,

$$g_i(\hat{c}-c) = B_i(\hat{c}-c)^{(1+\varepsilon_{\hat{c}-c})/\varepsilon_{\hat{c}-c}}\Big/\frac{1+\varepsilon_{\hat{c}-c}}{\varepsilon_{\hat{c}-c}},$$

where B_i is a firm-specific scale parameter. The firm maximizes after-tax profits accounting for the cost of evasion when facing a policy (τ, μ), which implies the following production and evasion choices:

$$y_i = A_i^{(1+\varepsilon_y)}(1-\tau_E)^{\varepsilon_y}\Big/\frac{\varepsilon_y}{1+\varepsilon_y}, \tag{A7}$$

$$c_i = F_i + A_i^{(1+\varepsilon_y)}(1-\tau_E)^{(1+\varepsilon_y)}, \tag{A8}$$

$$\hat{c}_i - c_i = \left(\frac{\tau\mu}{B_i}\right)^{\varepsilon_{\hat{c}-c}}. \tag{A9}$$

We assume a uniform production elasticity $\varepsilon_y = .5$. For each firm i, we calibrate the scale parameter A_i to match the turnover y_i reported on its tax return using (A7). We calibrate the parameters of the evasion cost function assuming a uniform evasion elasticity $\varepsilon_{\hat{c}-c}$ such that for each firm the evasion rate $(\hat{c}_i - c_i)/y_i$

equals .169 × (.005/.20) = .0042 and .669 × (.005/.35) = .0096 for $(\tau, \mu) = (.20, 1)$ and $(\tau, \mu) = (.35, 1)$, respectively, using (A7) and (A9). This corresponds to our evasion rate response estimates in table 2 for low-rate and high-rate firms, respectively, assuming a production elasticity $\varepsilon_y = .5$. Finally, for each firm i, we calibrate the fixed-cost parameter F_i such that the sum of the cost c_i and evasion $\hat{c}_i - c_i$ implied by the firm's optimal choices matches the cost reported on its tax return, using (A8) and (A9). We use the tax reports in 2008 when all firms were subject to a profit tax. We account for the preferential low rate that some firms face in the calibration but ignore this distinction in our numerical calculations for different policies.

For our numerical analysis, we calculate aggregate tax revenues and aggregate after-tax profits net of evasion costs for each tax policy (τ, μ). We use $(\tau, \mu) = (.35, 1)$ as the benchmark policy to determine the firms' aggregate profit constraint. When changing the assumed value for the production elasticity, we follow the same procedure to recalibrate our model such that it remains consistent with our empirical estimates and the distribution of reported turnover and costs.

D. *Additional Tables*

TABLE A1
Comparison of Missing and Nonmissing Observations

	Observations (1)	Median (2)	Mean (3)	Standard Deviation (4)
	A. Firms Reporting Profits and Turnover			
Profits	15,681	.1	−.7	12.3
Turnover	15,681	18.7	181.0	703.5
Salary	6,714	6.3	25.9	76.6
Interest	8,361	.5	10.7	49.9
Share of low-rate firms	15,681		.20	
	B. Firms Reporting Profits Only			
Profits	11,754	.0	−.8	12.5
Salary	2,800	9.1	35.9	94.5
Interest	4,185	.2	13.1	74.7
Share of low-rate firms	10,467		.16	
	C. Firms Reporting Turnover Only			
Turnover	8,546	9.2	454.5	7,401.3
Salary	3,073	5.0	37.9	272.0
Interest	3,763	.7	40.2	260.1
Share of low-rate firms	8,546		.27	

Note.—The table compares different samples of firms depending on whether or not they report profits and turnover. This is based on the final data (after consistency checks are applied) for all years pooled, excluding each year in the top and bottom 5 percent tails in terms of profits. Panel A considers firms that report both profits and turnover. Panel B considers firms that report profits only. Panel C considers firms that report turnover only. Columns 1–4 report the number of observations, median, mean, and standard deviation for different observable characteristics (turnover, profits, salary payments, interest payments, share of small firms). All statistics are in millions of Pakistani rupees.

TABLE A2
Data Cleaning Steps: Sample Definition

Sample	Definition
Firms reporting profits and turnover	Firms reporting profits Π, turnover y, and incorporation date D. Based on Π and y, derive implied tax liabilities $\tilde{T}^y$, $\tilde{T}^\Pi_H$, and $\tilde{T}^\Pi_L$ (high and low profit rate).
Consistency check I	Drop firm if reported and implied tax liability are inconsistent, i.e., $T^y \neq \tilde{T}^y$ or $T^\Pi \neq \tilde{T}^\Pi_H$ and $T^\Pi \neq \tilde{T}^\Pi_L$. If $T^\Pi = \tilde{T}^\Pi_H$ or $\tilde{T}^\Pi_L$, assign $\{H, L\}$. If T^Π is missing, assign $\{H, L\}$ based on y, D, and capital K.
Consistency check II	Drop firm if reported and implied taxpayer status are inconsistent, i.e., if $T^y > T^\Pi$ and $\tilde{T}^y < \tilde{T}^\Pi$; $T^y < T^\Pi$ and $\tilde{T}^y > \tilde{T}^\Pi$; $\tilde{T}^y > \tilde{T}^\Pi$ and T^y are missing; and $\tilde{T}^y < \tilde{T}^\Pi$ and T^Π are missing.

Note.—This table explains the consistency checks applied to the data. For all consistency checks, a tolerance threshold of 5 percent is used. Capital K is equity plus retained earnings. Note that the implied turnover tax liability used for consistency check I is gross implied turnover tax liability minus net deductions (which are deducted from the tax liability before the taxpayer status—turnover or profit taxpayer—is determined). For the same reason, the profits to turnover ratio used for consistency check II and for the bunching graphs is (profits minus net reductions)/turnover for firms that report positive net reductions.

TABLE A3
Data Cleaning Steps: Sample Size

Step and Year	High-Rate Firms	Low-Rate Firms
Raw data:		
2006/7/9	45,284	
2008	21,445	
2010	21,584	
Firms reporting profits and turnover:		
2006/7/9	10,228	2,899
2008	4,515	1,546
2010	4,862	1,867
After consistency check I:		
2006/7/9	10,260	2,197
2008	4,702	1,114
2010	5,193	1,415
After consistency check II:		
2006/7/9	9,467	1,965
2008	4,702	1,114
2010	4,661	1,238

Note.—This table displays the sample size for different steps in the cleaning process.

TABLE A4
Robustness of Bunching Estimates

	A. Varying the Order of Polynomial				
	3	4	5	6	7
High-rate firms, 2006/7/9	4.27	3.92	4.47	6.13	5.59
	(.1)	(.1)	(.1)	(.1)	(.1)
Low-rate firms, 2006/7/9	1.93	2.04	2.00	2.47	2.50
	(.2)	(.2)	(.2)	(.2)	(.2)
High-rate firms, 2010	2.53	2.23	2.04	1.48	1.41
	(.2)	(.2)	(.2)	(.1)	(.1)
	B. Varying the Number of Excluded Bins				
	1	2	3	4	5
High-rate firms, 2006/7/9	1.83	2.58	3.68	4.47	2.22
	(.1)	(.1)	(.1)	(.1)	(.1)
Low-rate firms, 2006/7/9	1.70	2.00	2.01	1.48	1.45
	(.1)	(.2)	(.2)	(.3)	(.3)
High-rate firms, 2010	1.81	2.04	2.55	2.41	2.31
	(.1)	(.2)	(.2)	(.2)	(.3)

Note.—The table presents estimates of the excess mass *b*, for different specifications of the estimating eq. (14), for the subsamples considered in table 2. Bunching *b* is the excess mass in the excluded range around the kink, in proportion to the average counterfactual density in the excluded range. Panel A presents estimates for different choices of the order of polynomial $q \in \{3, 4, 5, 6, 7\}$, for the excluded range chosen as in table 2 (four bins on either side of the kink for high-rate firms in 2006/7/9, two bins otherwise). Panel B presents estimates for different choices of the excluded range (one to five bins on either side of the kink) for the order of polynomial chosen as in table 2 ($q = 5$). Bootstrapped standard errors are shown in parentheses.

References

Allingham, Michael G., and Agnar Sandmo. 1972. “Income Tax Evasion: A Theoretical Analysis.” *J. Public Econ*. 1:323–38.

Andreoni, James, Brian Erard, and Jonathan Feinstein. 1998. “Tax Compliance.” *J. Econ. Literature* 36:818–60.

Auerbach, Alan J. 2002. “Taxation and Corporate Financial Policy.” In *Handbook of Public Economics*, vol. 3, edited by Alan J. Auerbach and Martin Feldstein, 1251–92. Amsterdam: Elsevier.

Auerbach, Alan J., Michael P. Devereux, and Helen Simpson. 2010. “Taxing Corporate Income.” In *Dimensions of Tax Design: The Mirrlees Review*, edited by James A. Mirrlees and Stuart Adam. Oxford: Oxford Univ. Press.

Baunsgaard, Thomas, and Michael Keen. 2010. “Tax Revenue and (or?) Trade Liberalization.” *J. Public Econ.* 94 (October): 563–77.

Besley, Timothy, and Torsten Persson. 2013. “Taxation and Development.” In *Handbook of Public Economics*, vol. 5, edited by Alan J. Auerbach, Raj Chetty, Martin Feldstein, and Emmanuel Saez. Amsterdam: Elsevier.

Burgstahler, David, and Ilia Dichev. 1997. “Earnings Management to Avoid Earnings Decreases and Losses.” *J. Accounting and Econ*. 24:99–126.

Cage, Julia, and Lucie Gadenne. 2014. “Tax Revenues, Development, and the Fiscal Cost of Trade Liberalization, 1792–2006.” Working paper, Univ. Warwick and Sciences Po Paris.

Carrillo, Paul, Dina Pomeranz, and Monica Singhal. 2015. "Dodging the Tax Man: Firm Misreporting and Limits to Tax Enforcement." Working paper, Harvard Univ.

Chetty, Raj. 2009. "Is the Taxable Income Elasticity Sufficient to Calculate Deadweight Loss? The Implications of Evasion and Avoidance." *American Econ. J.: Econ. Policy* 1 (August): 31–52.

Chetty, Raj, John Friedman, Tore Olsen, and Luigi Pistaferri. 2011. "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *Q.J.E.* 126:749–804.

Dharmapala, Dhammika, Joel Slemrod, and John Douglas Wilson. 2011. "Tax Policy and the Missing Middle: Optimal Tax Remittance with Firm-Level Administrative Costs." *J. Public Econ.* 72:1036–47.

Diamond, Peter A., and James A. Mirrlees. 1971. "Optimal Taxation and Public Production. I: Production Efficiency." *A.E.R.* 61 (March): 8–27.

Emran, M. Shahe, and Joseph E. Stiglitz. 2005. "On Selective Indirect Tax Reform in Developing Countries." *J. Public Econ.* 89:599–623.

Ernst & Young. 2013. *Worldwide Corporate Tax Guide*. Technical report. http://www.ey.com/Publication/vwLUAssets/Worldwide_corporate_tax_guide_2013/2013/$FILE/Worldwide_corporate_tax_guide_2013.pdf.

Federal Board of Revenue. 2013. *Quarterly Review*. Technical report (July–September). Pakistan: Fed. Board Revenue.

Federal Tax Ombudsman. 2013. "Federal Tax Ombudsman's Order in Review Appeal no. 12/2012, 2013." http://www.karachitaxbar. com/images/Newsletter01.10.2013.pdf.

Feldstein, Martin. 1999. "Tax Avoidance and the Deadweight Loss of the Income Tax." *Rev. Econ. and Statis.* 81 (November): 674–80.

Gordon, Roger, and Wei Li. 2009. "Tax Structures in Developing Countries: Many Puzzles and a Possible Explanation." *J. Public Econ.* 93 (August): 855–66.

Gruber, Jonathan, and Joshua Rauh. 2007. "How Elastic Is the Corporate Income Tax Base?" In *Taxing Corporate Income in the 21st Century*, edited by James R. Hines, Alan J. Auerbach, and Joel Slemrod. Cambridge: Cambridge Univ. Press.

Hassett, Kevin A., and R. Glenn Hubbard. 2002. "Tax Policy and Business Investment." In *Handbook of Public Economics*, vol. 3, edited by Alan J. Auerbach and Martin Feldstein, 1293–1343. Amsterdam: Elsevier.

Keen, Michael. 2008. "VAT, Tariffs, and Withholding: Border Taxes and Informality in Developing Countries." *J. Public Econ.* 92:1892–1906.

———. 2013. "Targeting, Cascading and Indirect Tax Design." Working Paper no. 13/57, Internat. Monetary Fund, Washington, DC.

Kleven, Henrik Jacobsen, Martin B. Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez. 2011. "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark." *Econometrica* 79 (3): 651–92.

Kleven, Henrik Jacobsen, Claus Thustrup Kreiner, and Emmanuel Saez. 2009. "Why Can Modern Governments Tax So Much? An Agency Model of Firms as Fiscal Intermediaries." Working Paper no. 15218, NBER, Cambridge, MA.

Kleven, Henrik Jacobsen, and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *Q.J.E.* 128:669–723.

Kopczuk, Wojciech, and Joel Slemrod. 2006. "Putting Firms into Optimal Tax Theory." *A.E.R. Papers and Proc.* 96 (May): 130–34.

Kumler, Todd, Eric Verhoogen, and Judith Frías. 2013. "Enlisting Employees in Improving Payroll-Tax Compliance: Evidence from Mexico." Working paper, Columbia Univ.

Mayshar, Joram. 1991. "Taxation with Costly Administration." *Scandinavian J. Econ*. 93:75–88.

Munk, Knud Jørgen. 1978. "Optimal Taxation and Pure Profit." *Scandinavian J. Econ*. 80:1–19.

———. 1980. "Optimal Taxation with Some Non-taxable Commodities." *Rev. Econ. Studies* 47:755–65.

Pomeranz, Dina. 2013. "No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax." Working paper, Harvard Bus. School.

Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Econ. J.: Econ. Policy* 2 (August): 180–212.

Saez, Emmanuel, Joel Slemrod, and Seth Giertz. 2012. "The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review." *J. Econ. Literature* 50:3–50.

Slemrod, Joel. 1995. "Income Creation or Income Shifting? Behavioral Responses to the Tax Reform Act of 1986." *A.E.R. Papers and Proc.* 85:175–80.

———. 2001. "A General Model of the Behavioral Response to Taxation." *Internat. Tax and Public Finance* 8:119–28.

Slemrod, Joel, Marsha Blumenthal, and Charles Christian. 2001. "Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota." *J. Public Econ.* 79 (March): 455–83.

Slemrod, Joel, and Caroline Weber. 2012. "Evidence of the Invisible: Toward a Credibility Revolution in the Empirical Analysis of Tax Evasion and the Informal Economy." *Internat. Tax and Public Finance* 19 (February): 25–53.

Slemrod, Joel, and Shlomo Yitzhaki. 2002. "Tax Avoidance, Evasion, and Administration." In *Handbook of Public Economics*, vol. 3, edited by Alan J. Auerbach and Martin Feldstein, 1423–70. Amsterdam: Elsevier.

Stiglitz, Joseph E., and Partha Dasgupta. 1971. "Differential Taxation, Public Goods, and Economic Efficiency." *Rev. Econ. Studies* 38:151–74.

Waseem, Mazhar. 2013. "Taxes, Informality and Income Shifting: Evidence from a Recent Pakistani Tax Reform." Working paper, London School Econ.

World Bank. 2009. *Pakistan Tax Policy Report: Tapping Tax Bases for Development.* Technical report. Washington, DC: World Bank.

Yitzhaki, Shlomo. 1974. "A Note on Income Tax Evasion: A Theoretical Analysis." *J. Public Econ*. 3:201–2.