Essays on Taxation in Limited Tax Capacity Environment

Mazhar Waseem

A thesis submitted to the Department of Economics of the London School of Economics and Political Science for the degree of Doctor of Philosophy in Economics, London, July 2013

Declaration

I certify that this thesis is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The second chapter draws on work carried out jointly with Michael Best, Anne Brockmeyer, Henrik Kleven and Johannes Spinnewijn. The Third chapter is based on a paper coauthored with Henrik Kleven. In addition to being involved in the genesis of these two project, my contribution to the two chapters includes researching contextual background and carrying out empirical analysis (for the 2nd chapter jointly with Anne Brockmeyer).

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 39,635 words.

Mazhar Waseem
July 2013

Abstract

I present three essays on income taxation in Pakistan. The first essay investigates how taxes influence agents' earnings, compliance and business organization choices. Using a tax reform introduced in Pakistan in 2010, which raised tax rates on partnership earnings as compared to sole proprietorship income, as a natural policy experiment, I (*i*) identify a full range of behavioral responses to the tax rate changes (*ii*) study the determinants of tax compliance (*iii*) investigate if VAT *causes* firms to be more tax compliant. Relying on administrative tax records that comprise the universe of income tax returns filed in 2006-11 and a rich set of firm characteristics, I find that the reform induced substantial extensive and intensive margin responses including reduction in earnings, income shifting, movement into informality, and spillover effects on VAT base. I also find that the firms that have greater fraction of their tax withheld at source, are registered for VAT, or withhold taxes of other agents are more tax compliant. This highlights the importance of the notion that information trails on arm-length business transactions facilitate enforcement. Comparing short-term responses of partnership firms – which arguably identify tax evasion – on both sides of the VAT exemption threshold, I find that the evasion changes discontinuously at the cutoff suggesting that the VAT causes firms to be more compliant. In the second essay, I along with my co-authors, analyze the design of tax systems under imperfect enforcement. A common policy in developing countries is to impose minimum tax schemes whereby firms are taxed either on profits or on turnover (with a much lower tax rate on turnover), depending on which tax liability is larger. This is a production inefficient tax policy, but has been motivated by the idea that turnover taxes are harder to evade. Such schemes give rise to kink points in firms' choice sets as the tax rate and tax base jump discontinuously at a profit rate threshold. Analyzing responses to one such scheme in Pakistan, we find large bunching of corporate firms around the minimum tax kink. We show that the combined tax rate and tax base change at the kink provides small real incentives for bunching, making the policy ideal for eliciting evasion. Based on the methodology that we develop, we estimate that turnover taxes reduce evasion by up to 60-70% of corporate income. In the third essay, I along with Henrik Kleven, develop a framework for non-parametrically identifying optimization frictions and structural elasticities using notches – discontinuities in the choice sets of agents – introduced for example by tax and transfer policies. We apply our framework to tax notches in personal income tax schedule of Pakistan to estimate structural elasticities of taxable income.

Acknowledgements

I am extremely grateful to my supervisor, Henrik Kleven, for his abiding support, guidance and encouragement over the years of this research. Without his help, three chapters of this thesis would not have been possible. The thesis also benefited a great deal from discussions with Michael Best, Anne Brockmeyer, Camille Landais, Adnan Khan, Matthew Skellern and Johannes Spinnewijn.

Finally, I am especially grateful to Sadia, Moazz, Fatima and Khadija for all of their love, understanding and support.

Contents

List of Figures

List of Tables

List of Appendices

Taxes, Informality and Income Shifting: Evidence from a Recent Pakistani Tax Reform

1.1 Introduction

One of the central challenges developing countries facing today is how to expand their capacity to raise taxes. It has been argued that buildup of public capital, just like private physical and human capital accumulation, is essential to the economic growth process (Besley & Persson 2013). One fundamental constraint preventing such buildup, however, is that personal income taxation and VAT, which raise bulk of government finances in industrialized countries, contribute so little in developing countries. Despite the importance of these taxes to the public finances of developing world, empirical evidence on their performance within such settings is limited.

In this paper, I use an income tax reform introduced in Pakistan in 2010 (hereafter "the reform"), which raised tax rates on partnership earnings as compared to sole proprietorship income, as a natural policy experiment to (*i*) identify a full range of behavioral responses to the tax rate changes (*ii*) study the determinants of tax compliance (*iii*) investigate if VAT *causes* firms to be more tax compliant. Before the reform, unincorporated businesses, which constitute more than 50% of the personal income tax filers in Pakistan, were treated symmetrically for the purposes of taxation. Their earnings were taxed through a graduated tax schedule comprising fourteen brackets with tax rates varying progressively from 0% to 25 %. The reform replaced this scheme with two different tax systems for partnership and sole proprietorship firms. For partnerships, a flat tax scheme involving a tax rate of 25%, with no exemption threshold, was introduced. For sole proprietorships, the graduated tax scheme was continued but number of brackets were reduced and bracket thresholds were moved such that a majority of sole proprietors experienced tax reduction.

Pakistan's tax administration, the Federal Board of Revenue (FBR), explained motivation for the reform in the following words: "In order to strengthen the drive for documentation, a uniform tax rate for small companies as well as AOPs[1] is proposed

[1] FBR uses the generic term Association of Persons (AOP) to denote partnership firms.

@ 25% of their taxable income"(FBR 2010). Purpose of the reform, hence, was to promote incorporation of partnership firms by making their taxation equal to small corporate firms. Unintentionally, however, it created a large tax rate variation between very similar firms in the unincorporated sector. The reform was announced on 06-06-2010 and officially took effect from 01-07-2010. It was, however, made retroactive for partnership earnings to the beginning of 2009.[2] Retroactive application creates additional variation across firms, which is exploited to identify behavioral responses to the reform that include reduction in earnings, income shifting, movement into informality, and spillover effects on VAT base.

For the empirical analysis, I use administrative data from FBR comprising the population of income tax returns filed in 2006-11 and a rich set of firm characteristics reported at the time of registration and updated from time to time. My main results are the following. I find substantial response to the tax rate changes along both extensive and intensive margins. The reform led to the exit or break up of a large number of partnership firms: the number of such firms reporting positive taxable earnings decreased by 41% in 2009, by another 27% in 2010, and by an additional 15% in 2011. Thus, within three years of the reform the number of tax paying partnerships had declined to 36% of the pre-reform level. The firms that did not exit, reported lower earnings. Compared to sole proprietorships, which did not experience tax rate changes in 2009 (control group), taxable earnings of the treated partnership firms declined by more than 50% (intensive margin elasticity of about 2.4). The response is cleanly identified, as both treatment and control group have identical pre-reform trends, which separate sharply at the time of reform.

Partnership and sole proprietorship business forms are close substitutes. A natural response to expect, hence, is that all or some of the individual owners of treated firms shift their earnings to sole proprietorship base. This could be a reporting effect or a real change in business organization. Exploiting longitudinal nature of the data, I track the owners of the partnership firms that exit and find that about 55% of such individuals report positive sole proprietorship earnings after the reform. This indicates that more than half of the extensive response observed at the firm level was because of income shifting. For owners of the partnership firms that do not exit, I separately estimate responses of all components of taxable income and find that around 46% of the reduced earnings of such individuals can be explained by income shifting to sole proprietorships.

Purpose of the reform, however, was to cause income shifting to corporate tax base. To see if the policy had the intended effect, I examine both the entry and stock of corporate firms. I find that though registration of new partnership firms declines by almost 50% after the reform, there is no discernible increase in entry of corporate firms. Also, the number of corporate firms reporting positive taxable income remains almost constant at

[2] Pakistani tax year runs from July to June; year t in this paper refers to the tax year from July t to June $t + 1$.

the pre-reform level suggesting that the reform did not meet its objectives, at least in the short-run.

Partnership firms, in most of the tax jurisdictions, are treated as *pass through* entities. This implies that the firm does not pay tax on its income; instead, the owners pay tax on their *distributive* share of the firm's taxable earnings. In Pakistan, however, partnership earnings are taxed at the firm level. This along with progressive taxation of sole proprietorship firms implies that individuals reporting partnership income experience a higher or lower tax liability as compared to if similar income is reported as sole proprietorship income.[3] I find that in pre-reform years partners experienced an average *partnership tax penalty* equivalent to about 4% of their taxable income, which increased to more than 15% after the reform. Partnerships are much more common in developing countries as compared to developed countries where they are restricted mostly to human capital intensive industries like accounting and law. Production complementarities, imperfect capital markets and agency problems arising from costly monitoring of employees are believed to be the main reasons for this ubiquity. The distribution and size of the partnership tax penalty reflects importance of these factors. More importantly, however, the penalty also reflects the significant welfare losses arising from the exit or break up of partnership firms.

As mentioned earlier, the reform had retroactive applicability. By the time the tax changes were announced, most of the *real* earning activity for the tax year 2009 had already taken place. Any systematic differences in filing or reported earnings between the firms affected and not affected by the reform will, therefore, identify tax evasion. Detecting non-compliance and investigating its determinants has traditionally been difficult because those engaging in it tend to keep the behavior concealed. A unique advantage of the context, however, is that it allows clean identification of non-compliance. Using the firm characteristics data, I investigate major policy and firm observables that influence compliance along both intensive and extensive margins. I find that larger firms evade less on both margins. Sophisticated firms, on the other hand, evade more on the intensive margin but are less likely to exit. I also explore importance of three mechanisms argued in literature to facilitate compliance – third party reporting, information trails on arm-length transactions, and whistle blowing (see Kleven *et al.* 2009; Slemrod & Kopczuk 2002 for theory and Kleven *et al.* 2011; Kumler *et al.* 2012; Pomeranz 2013 for evidence). Results confirm that these three mechanisms are significant in explaining observed compliance variation across firms.

Does VAT causes firms to be more tax compliant? This is a question of great academic interest and even greater policy importance, as one consistent advice to developing countries over the last two decades has been to introduce VAT. Firms registered for VAT

[3] This is conceptually similar to joint taxation of married couples in the US, which, given the progressive income taxation, gives rise to the "marriage penalty" or "marriage subsidy". See Eissa & Hoynes (2000) for details.

are linked to their suppliers and buyers through the invoice-credit mechanism built into VAT. Transactions between the linked firms generate paper trails reducing their ability to manipulate earnings or to leave the formal sector (see Pomeranz 2013; de Paula & Scheinkman 2010 for recent evidence). About 25% of the partnership firms affected by the reform were registered to remit VAT on their sales. By comparing non-compliance across firms registered and not registered for VAT, self-enforcing properties of VAT can be assessed. I find that VAT registered firms respond less along both intensive and extensive margins. Because VAT registration is not *randomly* assigned, such differences do not reflect causal effects. To draw causal inference, I compare firms in a narrow range on both sides of the VAT exemption threshold. As the compared firms are expected to be similar on all dimensions other than VAT registration, any heterogeneity in response would reflect the causal effects. I find that level of evasion changes discontinuously at the VAT exemption threshold suggesting that VAT does make the firms more tax compliant.

Rest of this paper is organized as follows. Section 1.2 develops conceptual framework, section 1.3 provides an overview of the context and describes data, section 1.4 discusses empirical methodology and results, and section 1.5 concludes.

1.2 Conceptual Framework

Focus of this paper is tax behavior of unincorporated firms – sole proprietorships and partnerships. These are worker-owned firms with no separate legal existence of their own. Profits of the firms are their owners' earnings and are taxed through the personal income tax system.[4] This implies that the standard utility maximization framework underlying the new tax responsiveness literature (please see Saez *et al.* 2012 for a recent survey of this literature) can be applied for analysis of the responses generated by the reform. In this section, I propose two extensions to the standard model to make it compatible with the Pakistani settings.

The standard model considers an individual's decision problem broadly as a choice between consumption c and multiple dimensions of labor supply captured by taxable income z. These dimensions include hours, effort, training, career choices and tax evasion/avoidance activities. Individuals are assumed to maximize utility $u(c, z)$ subject to a budget constraint $c = z - T(z) = (1 - \tau).z + E$, where $T(.)$ is tax liability, $\tau \equiv T'(.)$ is marginal tax rate, and $E \equiv \tau.z - T(z)$ is the virtual income generated by $T(.)$. Such maximization produces a taxable income supply function $z = z(1 - \tau, E)$, where optimal z depends on net-of-tax rate $1 - \tau$ and virtual income E. Assuming weak separability between consumption c and activities underlying z, Feldstein (1999) showed that elasticity of taxable income (ETI), $\varepsilon \equiv \frac{1-\tau}{z}.\frac{\partial z}{\partial(1-\tau)}$, is a sufficient statistic for deadweight loss of

[4] Pakistani law does not allow creation of limited liability partnerships. Therefore, owners of partnership firms remain liable for all obligations of the firm.

taxation arising from all margins of response including tax evasion.[5]

First extension to the standard model assumes that taxable income z could be of two types: the income earned as a sole proprietor z^s and the income earned as a partner in a partnership firm z^p. Each individual now maximizes a utility function of the form $u(c, z^s, z^p)$ and faces a budget constraint $c = z^s + z^p - T(z^s, z^p)$ where $T(z^s, z^p)$ is a potentially non-linear and non-separable income tax system through which z^s and z^p are taxed, $\tau^j \equiv \frac{\partial T}{\partial z^j}$ is the marginal tax rate on income of type j, and $E \equiv \sum_{j \in \{s,p\}} \tau^j . z^j - T(z^s, z^p)$ is the generalized virtual income. Utility maximization now generates two distinct income supply functions $z^j = z^j(1 - \tau^s, 1 - \tau^p, E), j \in \{s, p\}$. Since past ETI literature has not been able to find significant income effects,[6]I assume that optimal choice of z^j depends only on the two net-of-tax prices. This simple extension allows empirical specification to separately estimate two distinct responses to the tax changes: the own-price substitution effect captured by the intensive margin elasticity $\varepsilon^j_j \equiv \frac{1-\tau^j}{z^j} . \frac{\partial z^j}{\partial(1-\tau^j)}$ and the income shifting response captured by the cross-price shifting elasticity $\varepsilon^j_k \equiv \frac{1-\tau^j}{z^k} . \frac{\partial z^k}{\partial(1-\tau^j)}, j, k \in \{s, p\}$.

Generally, tax reforms are associated with discrete changes in tax liabilities and, hence, may trigger extensive margin responses as well. Increased tax burden could push agents on the margin of participation either to drop out of labor force or to move into informal sector. To consider such behavior, I incorporate a discrete participation choice into the model. Utility maximization now takes place over two stages. In the first stage, agents make optimal earning choices conditional on participation and in the second stage they decide whether to participate or not. Participation in formal sector, however, entails fixed utility gains q arising, for example, from warm glow or ability to use financial sector or better production technologies. An agent participates only if utility from participation $u(z^s + z^p - T(z^s, z^p), z^s, z^p) + q$ exceeds utility from non-participation, assumed to be u_0, that is iff $q \geq u_0 - u(z^s + z^p - T(z^s, z^p), z^s, z^p) \equiv \bar{q}$. Given a smooth distribution of q in the population represented by the distribution function $F(q)$, a fraction $\theta(\tau^s, \tau^p) \equiv 1 - F(\bar{q})$ of all agents participate. Any movements into and out of the formal sector owing to tax changes is captured by the participation elasticity $\eta \equiv \frac{1-t^\rho}{\theta(\tau^s,\tau^p)} . \frac{\partial \theta(\tau^s,\tau^p)}{\partial(1-t^\rho)}$, where t^ρ is the average tax rate on participation.

Individuals in this framework are alike and differ only in terms of their skills to produce income of each type. These skills will be reflected in the optimal earning choices z^s and z^p made by them. As mentioned earlier, partnerships are formed mainly to exploit production complementarities, overcome credit market constraints or reduce agency costs

[5] This, however, is subject to the caveat pointed out by Chetty (2009) that with tax evasion ETI is a sufficient statistic only if costs of evasion are pure resource costs rather than transfers to other agents (for example fines imposed on evaders).

[6] Gruber & Saez (2002) was the first study which considered both income and substitution effects in the context of income tax changes made in the US through a series of tax reforms in the 1980s. They found small and insignificant income effect. More recently, Kleven & Schultz (2011) analyzed behavioral responses of Danish income taxpayers over a period of 25 years and found statistically insignificant income effects for the self-employed.

of monitoring employees. Therefore, individual choosing $z^p > 0$ will search for and partner with entrepreneurs with whom they share such complementarities or who could help them overcome the market imperfection or agency costs. Thus, individual choices of z^p in this framework will be reflected at the aggregate level in the number of and the taxable earnings reported by the partnership firms.[7] Behavioral responses to changes in partnership income taxation, hence, could be studied by looking at the firm or the individual level outcomes. Indeed, one advantage of Pakistani context is that it allows identification of responses at both the firm and the individual level. Focusing on the individuals, however, has the added advantage that income shifting between z^p and z^s can be studied. Accordingly, in empirical section of the paper, I first consider responses of the partnership firms and then of the individuals.

1.3 Institutional Background and Data

Partnerships and sole proprietorship firms constitute more than 50% of all personal income tax filers in Pakistan. This section describes taxation regime of these firms and the administrative tax data I use for the empirical analysis.

1.3.1 The Tax Reform

Figure 1.1 plots the tax schedules applicable to these firms in 2006-11. Pre-reform schedule, which is common to both types of firms, is illustrated by the solid blue curve in the figure. It featured 14 tax brackets with fixed *average* tax rate – varying from 0% at the bottom to 25% at the top – applied to each bracket. In 2010, the schedule was replaced with two different tax systems. Red curve in the figure shows the flat tax scheme – comprising a tax rate of 25% and no exemption threshold – introduced for partnership firms with retroactive effect from the beginning of 2009. New tax system for sole proprietorship firms (gray curve in the figure) enacted three changes: it (*i*) reduced number of brackets from 14 to 5 (*ii*) raised bracket cutoffs but (*iii*) did not increase average tax rates. Movement of the bracket cutoffs, however, meant that sole proprietorships in certain areas of the income distribution experienced tax reduction. The most salient of such reductions was at the bottom of the income distribution, where exemption threshold was increased from Rs. 100,000 to Rs. 300,000 in 2010 and further to Rs. 350,000 in 2011. These changes led to huge tax differential between the two group of firms, especially at lower levels of income.

For two reasons, the reform generates almost ideal conditions for studying the effects of taxation on earnings and compliance choices of agents. First, it creates compelling quasi-

[7] My purpose here is not to describe a fully specified model of business organization choice. I focus only on the features that highlight the fundamental econometric issues in the empirical application.

experimental variation between very similar taxpayers. Both group of firms have similar initial earnings and tax rates but experience drastically different tax changes. Pre-reform earnings trends are parallel and diverge sharply for the treatment group at the time of reform. Tax changes introduced by the reform are also large, particularly at bottom of the earnings distribution where some of the partnership firms experience a tax rate hike of more than fifty times. Large and salient tax rate changes are especially useful in eliciting responses, as past work has shown that optimization frictions prevent taxpayers from reacting to small tax changes (Chetty *et al.* 2009; Chetty 2012; Chetty *et al.* 2011). Second, assignment to higher tax rates is based on business form and is not correlated with reported earnings. Identification, hence, will not be confounded by issues created by income based control groups such as mean reversion (please see Saez *et al.* 2012; Saez 2004; Slemrod 1998; Kopczuk 2012 for a detailed exposition of identification issues in the new tax responsiveness literature).

1.3.2 Partnership Tax Penalty

As mentioned earlier, partnership earnings in Pakistan are taxed at the firm level. This implies that the tax code is not neutral between the two forms of earnings studied in this paper. Individual reporting partnership income incur a higher or lower tax liability as compared to if similar income is reported as sole proprietorship income. To see this, consider a partnership firm j with taxable income Z_j and N_j number of partners such that $Z_j = \sum_{i=1}^{N_j} z_i^p$, where z_i^p is partner i's share of the firm's profits. This firm, for a tax system $T(z)$, incurs a tax liability of $t(Z_j).Z_j$, where $t(z) \equiv \frac{T(z)}{z}$ denotes the average tax rate applicable to earnings level z. The partners, however, face no further taxation on partnership income. In case they have earnings from other sources, the additional tax liability is calculated through the *averaging* method. For aggregate taxable income $z_i = z_i^p + z_i^s$ partner i pays an additional liability of $t(z_i).(z_i - z_i^p)$ where $t(z_i)$ is average tax rate applicable to z_i. Thus z_i^p attracts a tax liability of $t(Z_j).z_i^p$ if reported as partnership income and $t(z_i).z_i^p$ if reported as sole proprietorship income. For a non-linear tax system, $t(Z_j)$ and $t(z_i)$ are generally not equal, and individuals experience a partnership tax penalty of $z_i^p.[t(Z_j) - t(z_{i)}]$, which for a progressive tax schedule may be negative if $z_i > Z_j$.

To explore the importance and dynamics of the penalty, I plot in Figure 1.2 year-wise histograms of $z_i^p.[t(Z_j) - t(z_{i)}]/z_i$.[8] The variable represents additional average tax rate (in percentage points) that taxpayers experience on reporting z_i^p as partnership income rather than sole proprietorship income. Table 1.1 reports in column (2) the year-wise average values of the level of penalty in PKRs and in column (3) the year-wise average values of the

[8] While setting up the expression of partnership tax penalty, I ignore the possibility that individuals can be partners in more than one firm. In my empirical application, about 6% of all individuals reporting $z_i^p > 0$ are partners in multiple firms; I drop these individuals from the sample for Figure 1.2 and Table 1.1, because I observe only z_i^p and not its breakdown by firms.

penalty as a percentage of taxable income. Together, the evidence illustrates the following two points. First, prior to the reform a vast majority of taxpayers reporting $z_i^p > 0$ experience partnership tax penalty (about 91%) rather than subsidy. After the reform, tax rate on partnership income is never lower than the tax rate on sole proprietorship income, and hence everyone with positive partnership income experiences higher taxation. Second, though there is considerable variance, the partnership tax penalty is quite large and increases substantially after the reform. Average penalty is about 4 % of taxable income before the reform and increases by more than three times to 15% after the reform. Post-reform, the penalty is as high as 25% for a large number of taxpayer (42%). This happens because the reform increases tax rate on partnership income to a flat 25%, but makes sole proprietorship income up to 300,000 (350,000 in 2011) exempt from tax.

The partnership tax penalty captures in a simple way the willingness to pay (WTP) of individuals for partnership business form. The evolution of size and distribution of the penalty illustrates that such WTP is substantial. More importantly, however, the evidence also shows that incentives to misreport partnership income as sole proprietorship income exist and are strong even before the reform. The reform only enhances these incentives. This implies that the partnerships we observe prior to the reform comprise the individuals for whom productivity gains from making the co-labor arrangement formal are large enough to overcome the tax disincentives. Substantial increase in size of the penalty may lead to the exit or break up of such firms. This may involve a real change in business form or mere misreporting of income source. The pre-reform size and distribution of the penalty, however, shows that both cases will entail welfare loss for owners of such firms.

1.3.3 Data

For this study, I use two administrative datasets from FBR in Pakistan. To study behavioral responses to the reform, I use tax return data comprising the universe of income tax returns filed by corporations, unincorporated firms, self-employed individuals and wage earners in 2006-2011. The dataset has more than 5 million year-observations and contains variables corresponding to items reported on the tax return form. To study the major policy and firm observables that influence tax enforcement, I use registration data that includes a host of individual and firm characteristics reported at the time of registration and updated from time to time. Since July 2009, electronic return filing is mandatory for all partnership firms and for sole proprietorship firms only if they are registered for VAT or need to claim tax refund. More than 57% of the partnership returns and around 12% of the sole-proprietorship returns used in this study have been filed electronically. Rest of the returns were filed at designated bank branches and were fed into computers by an IT firm independent from FBR. FBR has been using this data for automated processing and payment of VAT and income tax refunds, which has ensured that the data is kept updated and free of errors.

1.4 Empirical Analysis

The reform creates tax rate variation for both sole proprietorship and partnership firms. In my empirical analysis, however, I focus only on partnerships and the individuals who own these firms. It is because (*i*) they experience the larger and broader of the tax variation and (*ii*) very effective control groups are available for this set of taxpayers.

1.4.1 Effects of the Reform on Partnership Firms

I begin the empirical analysis by examining the effects of the reform on partnership firms. First, I present graphical evidence on sharp changes in the number of and the taxable income reported by these firms. Later on, the response is decomposed to investigate intensive and extensive margin behavior separately.

1.4.1.1 Graphical Evidence

Figure 1.3 plots the taxable income distributions of partnership firms in 2006-11.[9] Panel A, which contains pre-reform distributions only, shows that the number of partnership firms filing for tax was increasing before the reform: the number increased by 9% in 2007 and 28% in 2008. Two other features of the pre-reform distributions are noteworthy: there is strong bunching of firms at notches in the pre-reform tax schedule and yearly distributions are remarkably similar to each other – though number of firms increase from year to year, the addition only expands the distribution vertically without any discernible change in shape. This suggests that earning and reporting decisions are strongly shaped and influenced by the tax system but for a given tax system they remain stable. To see effects of the reform, I contrast post-reform distributions with the 2008 distribution. Panel B of the figure makes such a comparison and illustrates the strong response generated by the tax rate changes. In contrast to the increasing pre-reform trend, number of tax paying partnerships decrease sharply after the reform by 41% in 2009, 27% in 2010 and 15% in 2011. Thus, within three years the number of partnership firms reporting positive taxable decrease to 36% of the pre-reform level. In addition to the large extensive response, the post-reform distributions also feature evidence of reduced reported earnings. Post-reform densities are higher at lower levels of income capturing a clear leftwards shift of the earnings distribution.

[9] Throughout the empirical analysis, I focus only on taxpayers with earnings between 0 and Rs. 650,000, which constitute more than 90% of the sample. I drop taxpayers in the rest of the income distribution because they experience the least of tax variation and density of tax filers there is too thin to estimate responses credibly.

The responses depicted in Figure 1.3 are concentrated in the earning range where the firms experience the largest tax increases. This suggests that the responses are driven by the tax changes. To draw causal inference, however, we need to look at behavior of similar firms not affected by the tax changes. As noted earlier, the reform creates a very *natural* control group – sole proprietorships firms for which tax system stayed the same in 2009. Earnings of sole-proprietorships, however, are reported by the owners in their personal income tax returns. Some of these individuals are also partners in partnership firms and face an incentive to shift earnings from partnerships to tax-favored sole proprietorships. To ensure that no one in the control group is affected by the tax changes, I include only those individuals in the group who drive their earnings exclusively from sole-proprietorship firms in all years considered in this study. Such individuals constitute a majority (more than 90%) of the self-employed in the sample.

Figure 1.4 shows the taxable income distributions of the control group in 2006-11. The comparison of the pre-reform distributions (Panels A of Figure 1.3 and 1.4) illustrates the similarity of the treatment and control groups. Both set of firms bunch at notches in the pre-reform tax system in an identical fashion indicating comparable behavioral responses to the tax system. Consideration of the 2009 distribution of the control group (Panel B of Figure 1.3) reveals that the reporting behavior of this group of firms does not change in the year: both the number of tax filers and taxable income evolve strictly in accordance with the 2006-08 trend. This confirms that the changes in behavior of the treatment group are tax-driven. Predictably, however, the 2010-11 distributions of the control group are different featuring strong response to the tax rate changes that become applicable to this set of firms from 2010.

Income Shifting

To explore the income shifting from partnerships to sole proprietorships, I plot in Figure 1.5 the breakdown of earnings reported by the owners of partnership firms. Panels A and B of the figure show the before and after analysis of partnership income (z^p) reported by these individuals. Expectedly, the distributions are qualitatively and quantitatively similar to the corresponding distributions of partnership firms and display comparable effects of the reform: the number of individuals with positive partnership income decrease substantially and the post-reform distributions shift leftwards suggesting intensive margin response. Panels C and D of the figure present density distributions of sole-proprietorship income (z^s) reported by the owners and provide clear evidence of the income shifting. After a very stable pre-reform trend, the number of individuals with positive sole proprietorship earnings increase and the distributions shift rightwards capturing the income shifting along intensive margin. To see the extent to which the income shifting compensates the loss of partnership earnings, in Panels E and F I plot aggregate taxable income $z = z^p + z^s$ reported by these individuals before and after the reform. The two plots demonstrate that the income shifting mitigates less than 50% of the loss of partnership earnings. In section 1.4.2, using differences-in-differences analysis, I decompose the shifting response into the two underlying intensive and extensive margin components.

The graphical evidence presented above provides very clear evidence of strong behavioral response to the reform by partnership firms. The density distributions, however, conflate a number of underlying margins of behavior, which need to be separated to identify structural parameters important for the tax policy.

1.4.1.2 Intensive Margin

The leftwards shift of the earning distribution of partnership firms (Panel B of Figure 1.3) provides clear evidence of intensive margin response to the tax rate changes. In this section, I use differences-in-differences (DD) methodology to estimate the elasticity underlying the response. Control group for the analysis comprises the individuals who drive their earnings exclusively from sole proprietorship firms in all the years 2006-11. Since the control group itself experiences tax changes in 2010, I restrict the period of estimation to 2006-09. In the tax responsiveness literature, DD research design has been implemented using both repeated cross-section and panel approaches. While repeated cross-section is considered more robust to *mean reversion*, panel approach is argued to be the right method if the *composition* of sample changes over time (See Saez *et al.* 2012; Saez 2004; Kopczuk 2012 for a detailed discussion on the merits and demerits of these approaches). As mean reversion is not likely to be a problem in the current context,[10] and the composition of sample does change owing to the large-scale extensive response, applying DD to a panel of firms is the most appropriate specification for this application. Accordingly, I estimate the following baseline model

$$\Delta ln(z_{it}) = \varepsilon.\Delta ln(1-\tau_{it}) + \alpha_0.1(i \in Partnership) + \alpha_1.1(t \in After) + \nu_{it}, \qquad (1.4.1)$$

where z_{it} is taxable income reported by firm i in period t, *Partnership* is an indicator for partnership firm, $After$ is a dummy for the post-reform year, and $ln(1-\tau_{is})$ is instrumented by the interaction term $1(i \in Partnership).1(t \in After)$.[11] The DD estimate of ε will consistently identify the elasticity of taxable income if it can be shown that *parallel trends* assumption holds – without the tax changes, reported earnings would have evolved

[10] Generally, taxpayers with above-mean (below-mean) income one year are expected to have lower (higher) earnings next year due to fluctuations of transitory component of earnings from year to year. This seriously obfuscates behavioral responses to taxation, especially if variation in tax rates between high and low income taxpayers is used as a source of identification. In the present context, however, there is no reason to expect that transitory income fluctuations will be correlated with business organization of taxpayers and, hence, will vary systematically across the treatment and control groups.

[11] As for a non-linear tax system τ_{it} changes endogenously with z_{it}, we need to instrument τ_{it} to ensure consistency of the estimated parameters.

identically for both the treatment and control groups.

To see if the assumption is satisfied and to provide non-parametric evidence on intensive margin response, Figure 1.6 plots the earnings growth path of the two group of firms. To explore heterogeneity across the earnings distribution, I present plots in three income ranges (0 250K], (0 450K] and (0 650K] for both unbalanced and balanced panel samples. The evidence in Panels A-F of the figure clearly demonstrates that the identifying assumption is satisfied for all the samples considered. Reported earnings trends are parallel prior to the reform but separate sharply at the time of the reform. The decline of reported earnings in the treatment group ranges between 0.48 and 0.83 log points across different samples, with low-income samples experiencing the highest decline.

Table 1.2 reports the taxable income elasticities estimated from the DD regressions. Column (1) of the table shows the income group, columns (2)-(4) the estimates from unbalanced panel regressions, and columns (5)-(7) the estimates from balanced panel samples. Balanced panel regressions are based on the firms that file in all the four years 2006-09 and report taxable income in the range indicated in column (1). Unbalanced panel estimates, on the other hand, are based on samples that include for year t the firms that file in both year t and $t+1$ and report taxable earnings in the range indicated in column (1). The results reflect the substantial behavioral response to the reform first seen in Figures 1.3 and 1.5. Elasticities for all the samples are large with point estimates ranging between 2.3 and 2.9. Two other features of the results are important: the elasticities decline gradually when higher income firms are added to the sample and the balanced panel estimates are slightly larger in comparison to the unbalanced panel estimates. Declining responsiveness along the income distributions captures the widely discussed correlation in literature between the size and formality of firms (see, for example, Kleven *et al.* 2009; Gordon & Li 2009; Kopczuk & Slemrod 2006). High-income firms are also the larger firms and have lesser ability to conceal their earnings. Firms in the balance panel regressions are different in the sense that they file consistently and in the process may have acquired greater awareness of tax laws or the ability to game the system. This is consistent with a similar result in Kleven & Waseem (2013) who find that consistent tax filers are more likely to bunch at tax notches and are less likely to make strictly dominated choices signaling their superior tax literacy.

As discussed in section 1.3.1, the tax increases on partnership firms were given retroactive effect. The reform was announced on 06-06-2010 but was made applicable to partnership earnings from 01-07-2009. By the time the reform was announced, most of the *real* earning activity corresponding to the tax year 2009 had already taken place. This implies that the dominant component of the response identified here is tax evasion.[12] The

[12] Some of the reduced earnings are due to misreporting of partnership income as sole proprietorship income. Such income shifting, however, is also in the nature of tax evasion as no real change of business form is involved.

settings, hence, offer a unique opportunity to explore important policy and firm level determinants of tax evasion. I report the results of this analysis in section 1.4.5 of the paper.

The elasticities presented in Table 1.2 are large as compared to those reported in past literature especially Kleven & Waseem (2013). This, however, should not be surprising because it is widely known in literature that ETI is not a structural parameter depending solely on underlying preferences and technologies. It, rather, is a function of tax system and, hence, may vary from reform to reform.[13] Specifically, large tax reforms, being salient and costly to ignore, generate larger responses (Chetty 2012; Chetty *et al.* 2011, 2009). Reforms targeted to narrow tax bases create opportunities of income shifting and also produce stronger responses. The tax rate changes instituted by the reform are both large and not very broad-based and accordingly induce strong behavioral responses.

1.4.1.3 Extensive Margin

Graphical evidence presented in section 1.4.1.1 shows that the reform triggered the exit of a large number of partnership firms. In this section, I use a three-step strategy to identify the elasticity governing the response. The strategy is visually illustrated in panels A-L of Figure 1.7. To be consistent with my earlier analysis, I focus only on the firms with positive earnings up to Rs. 650,000. The first step in the strategy is to estimate the counterfactual number of partnership firms in the post-reform years (the counterfactual). Panel A of the figure shows the evolution of tax filing in the treatment and control groups defined in section 1.4.1.1. The filing in the control group is on a weakly declining linear trend prior to the reform and continues to evolve on the trend even after the reform. In contrast to this, filing in the treatment group is increasing before the reform but declines sharply following the reform. Comparison of the two series suggests that one approach to estimate the counterfactual could be to use the conventional difference-in-difference specification augmented with separate linear time trends for the treatment and control groups. Given, however, that data on filing is available only for three pre-reform years, it is difficult to justify linear filing trend convincingly. A more conservative alternative to find the counterfactual is to assume that the number of partnerships in the post-reform years would have stayed at the pre-reform level. Evidence that this approach provides a lower bound on the response is presented in Panels B-D of the figure.

Pakistani tax code contains provisions that mandate taxpayers to file tax returns even if there is no *real* earning activity or taxable income from the activity is below the exemption threshold.[14] This implies that every year a number of taxpayers with zero earnings file tax

[13] In fact, Slemrod & Kopczuk (2002) have suggested that policy makers can optimally *choose* ETI by appropriately defining taxable bases.

[14] In general, all *registered* taxpayers are required to file tax returns. In certain cases, even unregistered taxpayers must file. Such situations include among other: if (i) a taxpayer had filed a return or paid tax in any of the preceding two year (ii) if a taxpayer owns a car, a house or certain other categories

returns (nil-filers). In Panel B of the figure, I plot two filing series for the treatment group: dark blue curve shows the firms with positive taxable earnings ($\leq 650k$), while light blue curve comprises both these and the nil-filer partnership firms. Comparison of the two series shows that the aggregate filing in the treatment group stops increasing in 2009 and becomes almost flat at the pre-reform level. Panel C, that plots the corresponding two series for the control group, demonstrates that filing in the control group evolves smoothly on the pre-reform trend. Panel D compares the aggregate filing in the treatment and control groups. Together, the evidence shows that absent the tax reform partnership firms with positive taxable income would at least have stayed at the pre-reform level. Both filing series shown in Panel B are affected by the reform. The aggregate filing series is affected because it does not feature non-filers[15]and the reduced entry because of the reform. However, the fact that this series stays at the pre-reform level implies that without the tax rate changes tax paying partnership firms also would not have decreased.

Panels D and F show the counterfactual filing series from the two alternative approaches. The counterfactual in Panel D is obtained by running a difference-in-difference regression on the two series in Panel A with separate linear time trend for the treatment and control groups. Counterfactual in panel F is the lower bound as discussed above. Comparison of the observed and counterfactual series shows that within three years the reform leads to at least 63% fewer partnership firms in the economy. This corresponds roughly to an extensive margin elasticity of around 3, as these firms experience an average decrease in net-of-tax rate of around 21%. The overall elasticity, however, masks considerable heterogeneity, as graphical evidence presented in section 1.4.1.1 illustrates that the response is not uniform throughout the earnings distribution.

To explore such heterogeneity, second step in the strategy is to estimate counterfactual *distribution* of partnership firms for the post-reform year. For lower bound exercise, 2008 distribution is the counterfactual distribution for all the post-reform years. To estimate counterfactual distributions for the linear trend approach, I use one important feature of the pre-reform distributions noted earlier: for an unchanged tax system, the distribution changes very little from year to year. Increase in the number of taxpayers only expands the distribution vertically without any discernible shift sideways. I, accordingly, find counterfactual distributions for the linear trend approach by shifting the 2008 distribution upwards proportionally to have the same mass as predicted by the DD counterfactual. These counterfactual distributions for the years 2009-11 are shown in Panels G, I and K of the figure respectively.

The observed and counterfactual distributions, however, are still not comparable as observed distributions feature large intensive margin responses. Since the earnings distri-

of immovable property.

[15] Enforcement of mandatory filing provisions is far from perfect – only about 1 million of the 3.5 million registered taxpayers file returns. Data shows that some of the firms that *should* have filed after the reform do not file, indicating that non-filing is as important a margin of extensive response as nil-filing.

bution shifts leftwards in 2009, the comparison of observed post-reform distributions with the counterfactuals may lead to underestimation of extensive response at lower levels of income and overestimation at higher levels. To make the two distributions comparable, last step in the strategy is to strip the observed distributions of intensive responses. Using the earning responses estimated in section 1.4.1.2, I impute the firms observed in post-reform years earnings they would have reported had there been no intensive response. Panels H, J and L of the figure show these distributions along with the counterfactual distributions. Counterfactual distributions illustrate the number of firms in various income bins had there been no reform-driven response at all; observed distributions stripped of intensive response show number of such firms had there been no response at the intensive margin. By comparing the two, extensive elasticities can be estimated throughout the income distribution.

These estimates are presented in Table 1.3. Column (1) of the table shows the income group, columns (2) and (3) the number of firms in the observed distribution stripped of intensive response and the counterfactual distribution respectively, and column (4) the extensive margin elasticities for 2009. Columns (5)-(10) contain corresponding estimates for the years 2010 and 2011. Results in Panels A and B are based on the two alternative approaches to estimate the counterfactual. Following conclusions emerge from the analysis. First, estimated elasticities are large capturing substantial extensive response to the reform. The response is precisely estimated: estimates from the two alternative approaches are pretty tight, especially for low-income firms. Second, the response decreases with earnings and increases over time. The dynamics of the response is also heterogeneous across income groups. Low-income firms respond almost immediately, while high-income firms respond over a longer horizon. Decreasing response with earnings is expected and captures heterogeneous returns from formality. Low-income firms have little productivity gains from operating in the formal sector and are always on the margins of participation and non-participation. The reform was a big earnings shock for such firms and pushed them into informality. Increasing response over time on one hand reflects the fact that the tax differential between partnership and sole-proprietorship earnings grows over time. Long-term responses are also larger because adjusting to a tax change takes time: individuals affected by the reform may need to find alternative occupations or business forms. This is particularly evident from the fact that high-income firms ($z > 400k$) that do not experience the increasing tax differential over time respond only over a longer horizon (their short-term responses are not significant). A detailed analysis on the causes and attributes of the firms that respond on the extensive margin is presented in section 1.4.5 of the paper.

1.4.2 Effects of the Reform on Partners

The conceptual framework presented in section 1.2 indicates that firm level outcomes reflect the choices made at the individual level. When a partnership firm responds on the intensive or extensive margin, it reflects the decisions of individual partners to leave the formal sector, reduce earnings, or shift income to other sources. In this section, I investigate the reporting behavior of partners to decompose the firm-level response into its underlying components.

1.4.2.1 Intensive Margin

Figure 1.8 plots the growth path of partnership income (z^p), sole-proprietorship income (z^s) and aggregate taxable income ($z = z^p + z^s$) in Panels A-C respectively. Treatment group for the analysis includes the individuals who report $z^p \neq 0$ in the pre-reform years 2006-08. Control group comprises the individuals who report $z^p = 0$ for all years in the sample. Since the control group itself experiences tax changes in 2010, the period of estimation is restricted to 2006-09. I focus on a balanced panel of taxpayers, because income shifting response can cleanly be identified by tracking the same individuals over time. Consistent with my earlier analysis, I consider the taxpayers with positive earnings up to Rs. 650,000. In Panel A of the figure, I compare the partnership earnings reported by the treatment group to the aggregate taxable income reported by the control group. The plot features the steep decline of partnership earnings at the time of reform. Panel B of the figure explores income shifting[16] and illustrates that the sole-proprietorship earnings reported by the treated individuals increase sharply in comparison to the control group. The income shifting is responsible for relatively smaller response of aggregate taxable income reported in Panel C of the figure.

Plots in the figure also report the elasticity estimates from the following regressions

$$\Delta ln(z^k_{it}) = \varepsilon^k.\Delta ln(1 - \tau^p_{it}) + \alpha_0.1(i \in Partner) + \alpha_1.1(t \in After) + \nu_{it}, \qquad (1.4.2)$$

where z^k_{it} is the income of type k reported by individual i in period t, ε^k is the elasticity of income type k with respect to $1 - \tau^p$, $Partner$ is an indicator for treatment, $After$ is a dummy for the post-reform year, and as earlier $ln(1 - \tau^p_{it})$ is instrumented by the interaction term $1(i \in Partner).1(t \in After)$.[17] One difficulty with estimating elasticities from (1.4.2), however, is that partnership firms do not report the breakdown of their earnings by individual partners. This makes it difficult to determine exactly the pre-reform marginal tax rate experienced by individuals on partnership earnings reported by

[16] Only a few of the treated individuals shift earnings to wage or capital income. I focus only on income shifting between z^p and z^s because it is the predominant margin of response.

[17] Pre-reform earnings growth trends are not parallel for Panel A of the figure. To control for this, specification for elasticity reported in Panel A also includes separate linear time trend for the treatment and control groups.

them (post-reform rate is flat at 25%). To get around this difficulty, I assume that all partnerships comprise two partners who divide the firm's earnings equally. Since it is a very conservative assumption,[18] it very likely provides lower bound on the elasticities.

These elasticities are shown in respective panels of Figure 1.8. Partnership income elasticity (point estimate 0.95) though quite large is less than half of the corresponding firm-level elasticity (nearly 2.4). Apart from the fact that individual-level elasticity is a lower bound, the elasticity is smaller also because firm-level intensive margin response captures both intensive and extensive margin behavior at the individual level. Reduction in reported earnings of a partnership firm could also come from some of the partners leaving the firm and the formal sector. Such response is not captured in Panel A of the figure or in the elasticity reported. Cross-price elasticity of -0.85 (Panel B) indicates that taxpayers shift earnings considerably after the reform. Overall elasticity of 0.51, which is a weighted average of the two elasticities,[19] suggests that despite significant income shifting the reform resulted in substantial reduction of overall earnings reported by the treated individuals.

1.4.2.2 Extensive Margin

Figure 1.9 investigates the extensive margin responses of partners. The definitions of the treatment and control groups are the same as in the last section (1.4.2.1). To explore the importance of income shifting, two margins of extensive response are considered separately. Pure extensive response comprises the treated individuals who stop filing after the reform (non-filers) or file but report zero earnings (nil-filers). Switching response consists of the treated individuals who report zero partnership income but positive overall earnings after the reform.[20] To study switching, it is important that the same individuals are observed over time to see how the composition of reported income responds to the reform. This suggests the use of a balanced panel of individuals who file in all the years 2006-11. With balanced panel, however, pure extensive response cannot cover the non-filers. Given that income shifting margin is more important at the individual level, I use a balanced panel for the analysis with the tradeoff that the extensive response identified here covers nil-filers only.

Panel A of the figure illustrates the log number of filers in the treatment and control groups in 2006-11. For the control group, the series evolves smoothly and shows no signs of break at the time of reform. For the treatment group, the number of individuals

[18] Given the progressive tax schedule, percentage change in net of tax rate is the highest under this assumption. Under any alternative assumption (for example three partners dividing the firm's earnings equally), we will be attributing a given change in income to a lower percentage decrease in net of tax rate and will be estimating higher elasticities.

[19] The weights are a function of pre-reform levels of partnership and sole-proprietorship earnings and the percentage change in net of tax rate experienced by the individuals.

[20] In this section, switching to sole proprietorship business form only is considered. Next section explores switching to corporate business organization.

reporting positive earnings drops significantly at the time of reform. In Panel B of the figure, I plot the filing series for the treatment group along with a counterfactual obtained from running a difference-in-difference regression on the two series in Panel A. To account for the fact that filing in the control group is also affected by the tax rate changes that become applicable in 2010, I estimate the DD for the period 2006-09 only and extrapolate the counterfactual to 2010-11. The comparison of the observed and the counterfactual series shows that by 2011 the reform caused the exit of about 28% of the treated taxpayers. As the treated individuals experience a participation tax rate change of roughly 20%, this corresponds to an extensive margin elasticity of more than one on account of nil-filing only.

To explore switching, I repeat the above steps but treat the individuals who report zero partnership earnings but positive overall earnings as nil-filers. Expectedly, the filing in the treatment group (blue curve in Panel C) shows larger effects of the reform. Compared to the DD counterfactual (Panel D), the observed tax filers are now fewer by 43%, 54% and 63% in 2009, 2010 and 2011 respectively. This suggests that the reform induced 63% of the individuals who were partners for at least three consecutive years in partnership firms to quit the firms: about 45% of them disappeared into informality, the rest switched business organization to sole proprietorships.

1.4.3 Income Shifting to Corporate Tax Base

As explained in section 1.3.1, the purpose of the reform was to promote incorporation of partnership firms by bringing their taxation on a par with corporate firms. In this section, I investigate if the policy was able to achieve its objective. Theoretically, a firm's decision to incorporate is influenced by a variety of factors. Incorporation offers limited liability,[21] legal continuity and perpetual existence. Corporations, however, are costly to create and maintain. They need to keep audited accounts, face higher regulations, and experience double taxation. While making organizational form choice, entrepreneurs tradeoff these costs and benefits. The degree to which this decision is influenced by tax incentives is not clear, especially in a developing country settings. Past empirical literature on the subject, mainly based in the US, has found small to moderate effects (Gordon & MacKie-Mason 1994, 1997; Goolsbee 1998, 2004).

To evaluate if the reform significantly influenced the incorporation choice of firms, I examine in Figure 1.10 both the *entry* into and the *stock* of corporate sector in Pakistan. Potentially, the reform can spurt the entry of new corporations through two different channels. Some of the existing partnership firms may incorporate if the adjustment costs of doing so do not exceed the returns. Also, some new firms which absent the tax changes would have entered as partnerships might enter as corporations. Panel A of the figure

[21] As noted earlier, Pakistani law does not allow creation of limited liability partnerships, which are permitted in many other tax jurisdictions.

shows the month-wise registration of new firms with FBR in 2006-12. The series for corporate firms shows no signs of structural break at the time of announcement of the reform.[22] Though, there are considerable fluctuations, the number of new registrations settles to an almost constant level six months prior to the reform and continue to evolve on the trend after the reform. Contrary to this, the series for partnership firms shows clear signs of tax-driven response: the entry of new firms declines by almost 50% after the reform.

This evidence of weak or no effect is further strengthened by examining the stock of corporate firms. Panels B of the figure plots the taxable income distributions of corporate firms in Pakistan in 2006-11. The yearly histograms show no discernible changes over the years, and the post-reform empirical distributions[23] are extremely similar to the pre-reform distributions. We can contrast these to the corresponding distributions of partnership firms (Figure 1.3) to rule out any response.

The evidence, hence, suggests that the reform had no short-run effects on the incorporation choice of firms. The result, however, needs to be careful interpreted. Incorporation is a complex decision involving non-trivial adjustment costs. Appropriate time frame for evaluating the response, hence, is medium to long term when all firms are expected to have adjusted to the new incentives.

1.4.4 Spillover Effects on VAT Base

About 25% of the partnership firms that reported positive taxable income in 2008 were also registered to remit VAT on their sales. Changes in their behavior will affect VAT collections as well. The firms that respond on the intensive margin will remit lower VAT owing to the reduction in taxable base. The firms that exit the formal sector will be lost to VAT as well. The firms that break up into smaller sole proprietorships may also be lost if the new firms fall below the VAT exemption threshold.[24] In this section, I investigate these spillover effects generated by the reform.

In Figure 1.11, I plot the taxable income distributions of firms stratified by their VAT registration status. Sample for the analysis and the definitions of the treatment and control groups are exactly identical to subsection 1.4.1. The 2006-07 distributions are not shown for space considerations, but changes in the number of firms in these years are indicated in respective plots. The comparison of Panels A and B shows that the VAT registered partnership firms also respond to the reform though their responses are considerably smaller as compared to the rest of firms. The corresponding distributions

[22] If a new corporation is created by incorporation of an existing partnership firm, it will also show up as a new registrant in this series as corporations are required to register separately.

[23] Here 2010 and 2011 distributions only are *treated*, because taxpayer learn the tax rate changes on 14-06-2010.

[24] Manufacturing and retail firms with annual sales of not more than Rs. 5 million are exempt from payment of VAT.

for the control group in Panels C and D of the figure show that the responses are driven by the tax changes.

To quantify the effects, I use the methodology presented in subsections 1.4.1.2 and 1.4.1.3 to separate the intensive and extensive margin behavior. Figure 1.12 compares the earnings growth of VAT-registered firms with the rest of firms. Table 1.4 reports the intensive margin elasticities estimated from the difference-in-difference specification (1.4.1) using a balanced panel of firms. Column (7) of the table shows that intensive elasticities for VAT-liable firms are significant, more than one for all the sub-samples.[25] They are, however, smaller (about half) as compared to the firms not registered for VAT.

Figure 1.13 shows the breakdown of the aggregate extensive response, seen earlier in subsection 1.4.1.3, by VAT liability of firms. Here, I restrict the analysis to the lower bound exercise only. Panel B of the figure illustrates that the reform caused large-scale exit of VAT remitting partnership firms: within three years the number of such firms was down to 56% of the pre-reform level.[26] The comparison of Panels B and D, however, shows that even on the extensive margin VAT-registered firms respond less in comparison to other firms.

The smaller response of VAT-liable firms is quite consistent with similar evidence in past literature. It is claimed that such firms are linked to their suppliers and buyers through the invoice-credit mechanism built into VAT. Being part of such a chain creates a paper trail and reduces a firm's ability to manipulate its earnings or to leave the formal sector (see Pomeranz 2013; de Paula & Scheinkman 2010 for recent evidence). However, one difficulty with drawing causal inference in the present context is that VAT registration is not *randomly* assigned. Smaller responses may reflect other characteristics of such firms correlated with compliance, for example size. In section 1.4.6, I control for these confounding factors to explore the causal effects of VAT registration on the tax compliance of firms.

1.4.5 Which Firms Respond to the Reform?

Firms remit more than 80% of the tax collected in advanced countries (Christensen *et al.* 2001; Shaw *et al.* 2010). For developing countries, this percentage could be even higher though no reliable estimates are available. It is, hence, not surprising that investigating the determinants of firm compliance has been an area of great importance, especially for

[25] These elasticities are indicative of but not the exact measure of losses in VAT base. VAT base, under certain conditions, is equal to profits plus wages. These elasticities capture the effects operating through the margin of profits only. Potentially, wages paid by the firms may also respond to the tax rate changes, in which case the elasticities shall not reflect the losses exactly.

[26] Partnership firms constitute only a small fraction (about 4%) of the total VAT base. Also VAT collections are very skewed and more than 90% of the tax is remitted by top 300 firms that are either corporations or are not affected by the reform because of having earnings greater than Rs. 1.3 million. Thus, the exit of treated partnership firms will have little short-run effect on overall VAT collection in the country.

the countries with low tax capacity. Still, however, the empirical evidence in existing literature is limited owing to the obvious difficulty that tax evasion is hard to observe. One unique advantage of the present context is that the noncompliance of firms is cleanly identified. In this section I use the evidence to analyze the factors underlying a firm's decision to comply with tax laws.

1.4.5.1 Intensive Margin

I have argued earlier (subsection 1.4.1.2) that, owing to the retroactive applicability, the intensive response to the reform mostly identifies tax evasion . In Table 1.5, I investigate the relationship between tax evasion and firm observables by including triple interaction terms in (1.4.1). I omit the tax variable from the specification, so that the coefficient on interaction term $1(i \in Partnership).1(t \in After)$ captures the log change in reported earnings of the treated firms in the post-reform year. The details of the additional variables used in the regressions are provided in Appendix **??**. Results in columns (2)-(4) of the table show that while larger and more experienced firms evade less, sophisticated firms evade more. Columns (5)-(7) of the table explore the importance of the three mechanisms that have been argued in literature to improve compliance. Third party reporting provides independent information on taxable base to the government, and hence reduces a firm's ability to under-report earnings (Kleven *et al.* 2011, 2009). Firms registered for VAT are linked to their suppliers and buyers through the invoice-credit mechanism. This generates a paper trail of taxable transactions which can be observed more easily (Kopczuk & Slemrod 2006). Firms that act as withholding agents face greater risks of being caught because of whistle-blowing possibilities (Kleven *et al.* 2009; Kumler *et al.* 2012). Results confirm that all these three mechanisms are important determinants of compliance. In fact, column (9) of the table illustrates that reported earnings of a firm that has all the six attributes used as interactions increased rather than decreased after the reform.

1.4.5.2 Extensive Margin

We saw in section 1.4.1.3 that the reform caused the exit of a large number of partnership firms. To explore the major determinants of the response, I report in Table 1.6 results from the following regression

$$Exit_{it} = \beta_0 + \beta_1.X_i + \beta_2.Z_i + \nu_{it}, \tag{1.4.3}$$

where $Exit_{it}$ is an indicator that firm i reported positive earnings in 2008 but did not in year t, X_i are the firm observables of interest, and Z_i are a rich set of controls containing (i) tax office fixed effects (15 categories that capture the location and type of tax office)[27] (*ii*) industry fixed effects (six-digit industry code the firm belongs to) and (*iii*) age fixed effects (10 categories). All variables are introduced into the regression specification non-parametrically so that there are a total of 339 dummy variables in equation (1.4.3). Consistent with my earlier analysis, I focus only on the firm observables that have been emphasized in the tax enforcement literature to be important determinants of formality. Results in columns (2)-(5) show that the firms that are (*i*) large (*ii*) sophisticated (*iii*) have greater fraction of their tax withheld at source (*iv*) registered for VAT or (*v*) withhold the tax of other agents are less likely to exit as compared to other firms in the same industry, tax office and age decile.

The above analysis is, however, subject to few caveats. One disadvantage of using so many controls is that I lose observations where data is missing on any of the variables in the regression equation (results in Table 1.6 are based on around 50% of the potential sample). This could bias the results, if the missing observations are systematically different on important dimensions studied here. However, columns (1)-(3) of the table show that exit probability for the included firms is 41%, 59% and 63% in the years 2009, 2010 and 2011 respectively, which is not different from such probabilities for the complete sample (Figure 1.7, Panel F). This shows that the missing observations are not different in terms of the dependent variable. To further allay the concern, I drop from the regression equation the controls with the most missing data (industry code and annual sales). Such specifications produce results (not shown) very similar to those reported in the table. One additional worry with the results is that the regression specification treats every firm that leaves after 2008 as a tax-driven response. Some of the firms that exit would have done so regardless of the tax rate changes. To allay this concern, I re-estimate (1.4.3) with a sample that comprises only the firms that report positive taxable income in all the three pre-reform years 2006-08. These are regular filers the exit of which is more likely to have been driven by the tax rate changes only. These regressions also produce qualitatively very comparable results.

1.4.6 Is VAT Causal?

The evidence in sections 1.4.4 and 1.4.5 shows that VAT registered firms respond less along both intensive and extensive margins. The result holds even when a rich set of control are introduced to remove the influence of other confounding factors. In this section, I provide additional evidence to see if the observed relationship is causal.

The manufacturing firms with sales up to Rs. 5 million are exempt from remitting VAT on their sales. In Panel A of Figure 1.14, I compare the intensive margin responses, which arguably identify tax evasion, of partnership firms on both sides of the cutoff in the sales

[27] Pakistan has two types of tax offices. Large Taxpayer Units (LTUs), located in Karachi, Lahore and Islamabad, cater to top tax contributors. Regional Tax Offices (RTOs), located in twelve cities, administer the rest of firms.

distribution. Since the compared firms are expected to be similar on all dimensions other than VAT, any heterogeneity in behavior will capture the causal effects of VAT. I leave the firms with sales within a window of Rs. 0.5 million across both sides of the cutoff owing to the concern that their responses might be influenced by the threshold. The plot illustrates that the average earnings decline of the firms with sales $\in$ (3.5m 4.5m] is more than 0.2 log points larger in comparison to the firms with sales $\in$ (5.5m 6.5m]. To rule out the possibility that the difference is driven by the size variation, I compare in Panel B the firms with sales $\in$ (2.5m 3.5m] to the firms with sales $\in$ (3.5m 4.5m]. For these two set of firms the response is entirely homogeneous. The discontinuous change in the behavior of firms at the VAT exemption threshold, thus, suggests that registering for VAT makes the firm more tax compliant.

The above analysis is based on manufacturing firms, which comprise around 60% of the population of partnership firms. Some of these firms (around 33%) do not report their sales in income tax returns. Also, to be consistent with my earlier analysis, I drop the firms with taxable income of more than Rs. 650,000. This along with the skewed size distributions reduces the sample size around the VAT cutoff considerably. This has two implications for the analysis. First, I cannot carryout the sensitivity analysis of the result by varying the size of window around the cutoff. Reducing the window size will strengthen the claim that the compared firms are similar, but this leaves too small a sample to make any meaningful inference. Second, I leave out firms with earnings less than Rs. 30,000 in the years prior to the reform because they are subject to serious mean-reversion problem. I have argued earlier that mean reversion is not a problem for my application and accordingly have not dropped any observations on this account for the analysis in previous sections. This is because the tax variation exploited in the earlier sections was based on business form and was not correlated with the earnings of the compared agents. Here, however, I utilize differences in the behavior of firms in different parts of the sales distribution to make the inference, and as sales are strongly correlated with earnings, mean reversion is likely to confound the tax-driven responses. Dropping such firms from the sample is based on the suggestion in past literature that taxpayers at the bottom of the income distribution seriously aggravate the mean reversion problem (for example, Gruber & Saez 2002 drop taxpayers whose income is below $10,000 in the base year).

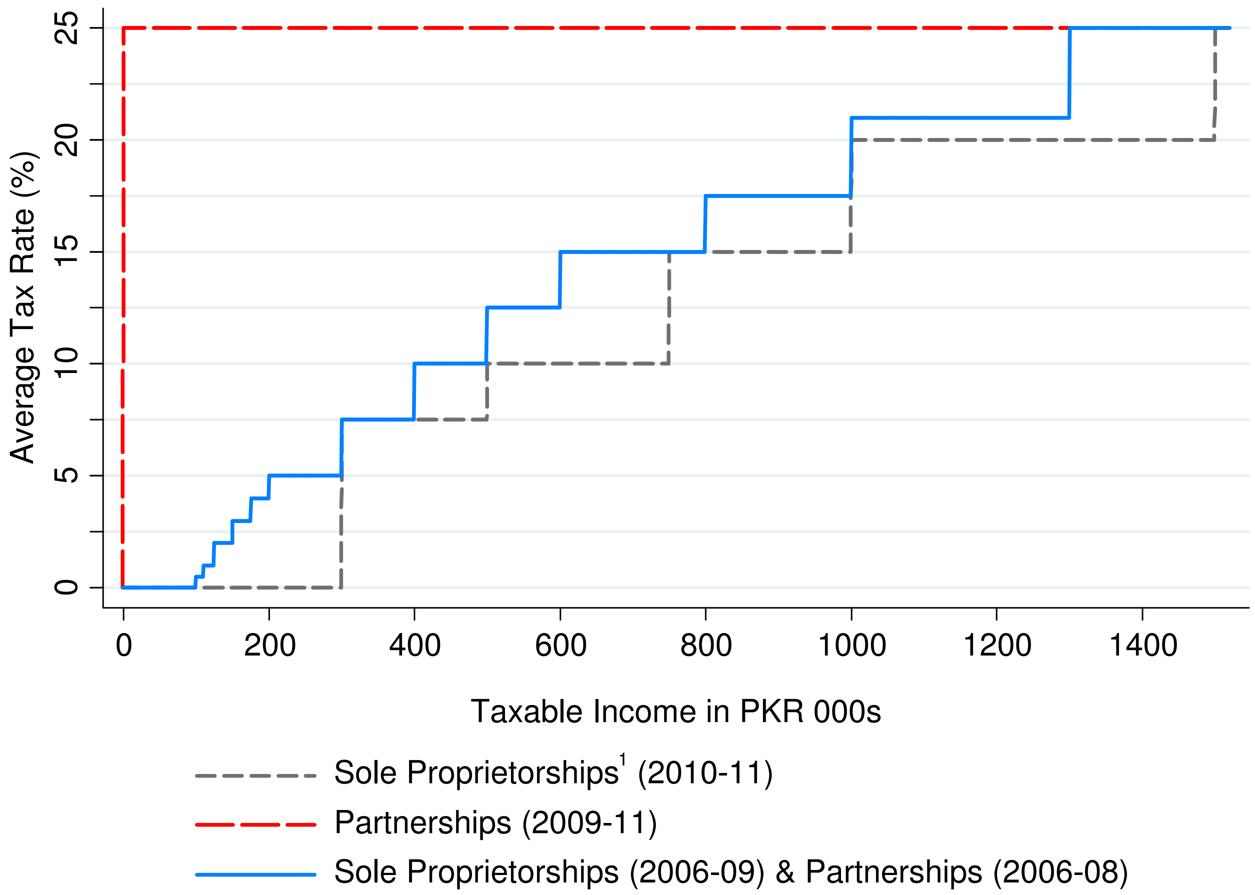
1.5 Conclusions

This paper has analyzed the effects of personal income taxation on the earnings, compliance and business organization choices of agents in a limited tax capacity setting. The 2010 reform induced the breakup or exit of a large number of partnership firms. The surviving firms reported substantially lower earnings. The finding that the tax base is quite elastic, especially at the bottom of the income distribution, has three important lessons for tax policy in developing countries. First, the tax system should actively seek to mitigate the costs of operating in the formal sector by instituting progressively increasing tax rates. Low-income firms are particularly sensitive on the participation margin and bringing their taxation on a par with high-income firm through flat rate schemes is likely to lead them to informality. Declining intensive margin responsiveness along the income distribution further strengthens the case for progressive taxation. Second, large tax rate changes are known to produce large behavioral responses and should be avoided to protect tax bases and limit efficiency costs of raising taxes. Third, tax rate increases on a narrowly defined tax base produce large incentives and opportunities for income shifting. Such reforms are less likely to produce additional revenues but impose significant welfare costs. Tax rate changes, hence, should be as broad-based as possible.

The retroactive applicability of the reform creates additional variation across firms that has been exploited to uncover non-compliance. The estimates of short-run responses to the reform provide perhaps the cleanest evidence in literature on how firm and policy variables influence tax evasion. Results show that the firms that have greater fraction of their tax withheld at source, are registered for VAT, or withhold tax of other agents are more tax compliant. These results generally support the notion that information trails on arm-length business transactions facilitate enforcement. The governments in the developing world, hence, need to make greater use of the tax instruments that promote compliance by generating information flows. Tax withholding on specific business transaction, however, leads to different effective tax rates across firms. Optimality of such schemes, hence, needs to be carefully evaluated on the basis of the tradeoff they offer between lower enforcement costs and production inefficiency.

Figure 1.1

The Tax Reform

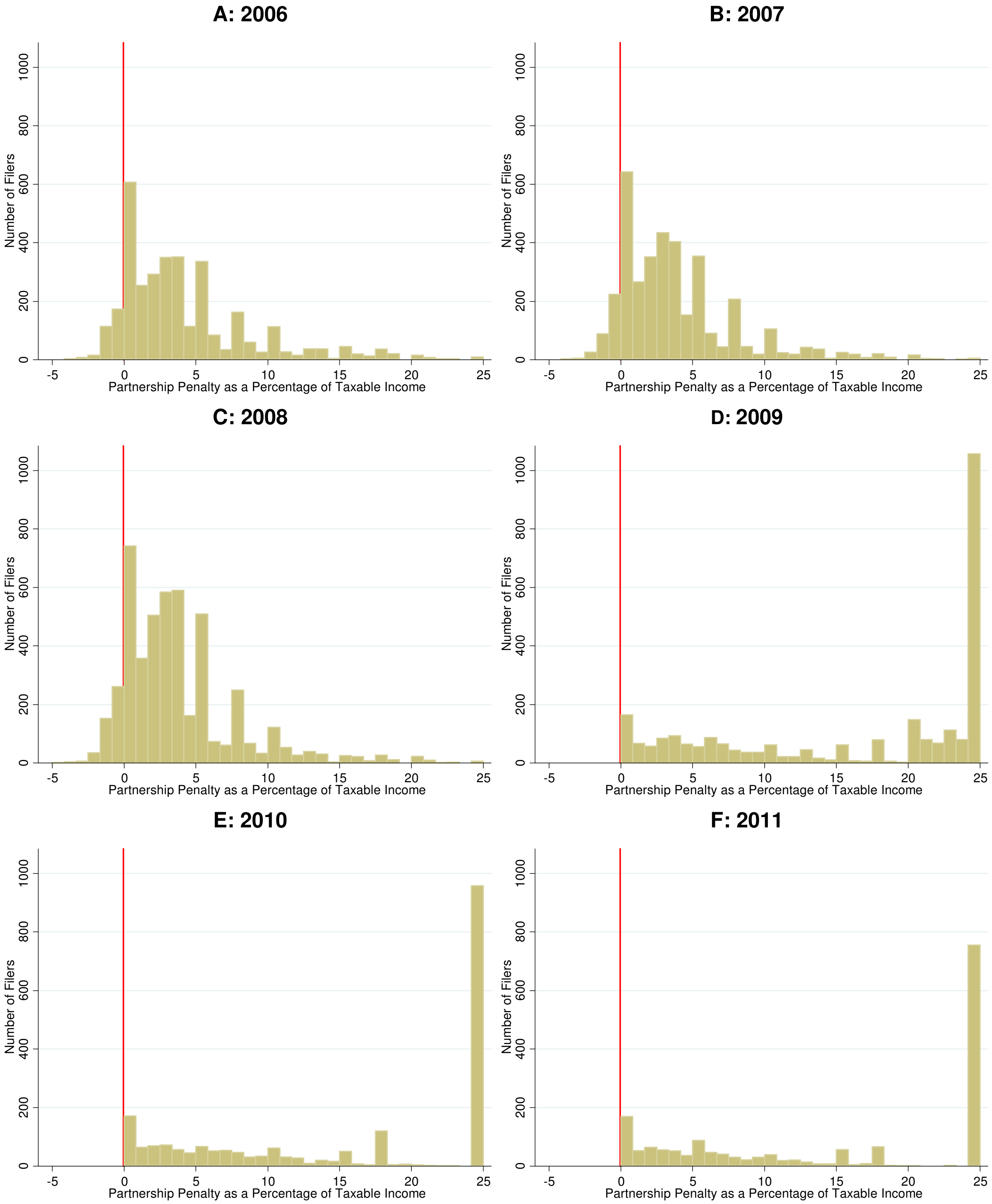


[1] Exemption threshold for self-employed individuals for the year 2011 is PKR 350,000.

Notes: the figure shows the Pakistani income tax regime for unincorporated businesses in 2006-11. Solid blue curve plots the tax schedule applicable to both sole proprietorships and partnerships in 2006-08. It features fourteen brackets with fixed *average* tax rate – varying from 0 to 25% – applied to each bracket. Post-reform schedules for partnership income is depicted by dashed red curve. It is a flat rate structure of 25% with no exemption threshold. The reform was introduced on 14-06-2010 but was made retroactive on partnership earnings from 01-07-2009. Post-reform tax system for sole-proprietorship earnings is shown by dashed grey curve. It consists of six brackets with average tax rate varying from 0% to 25%. Exemption threshold for the sole-proprietorship income was increased from PKR 300,000 to PKR 350,000 in 2011. All schedules show variations in average tax rates as a function of annual taxable income. Brackets' boundaries where tax rate changes are included in lower tax brackets. Taxable income is shown in thousands of Pakistani Rupees (PKR), and the PKR-USD exchange rate is about 100 as of July 2013

Figure 1.2

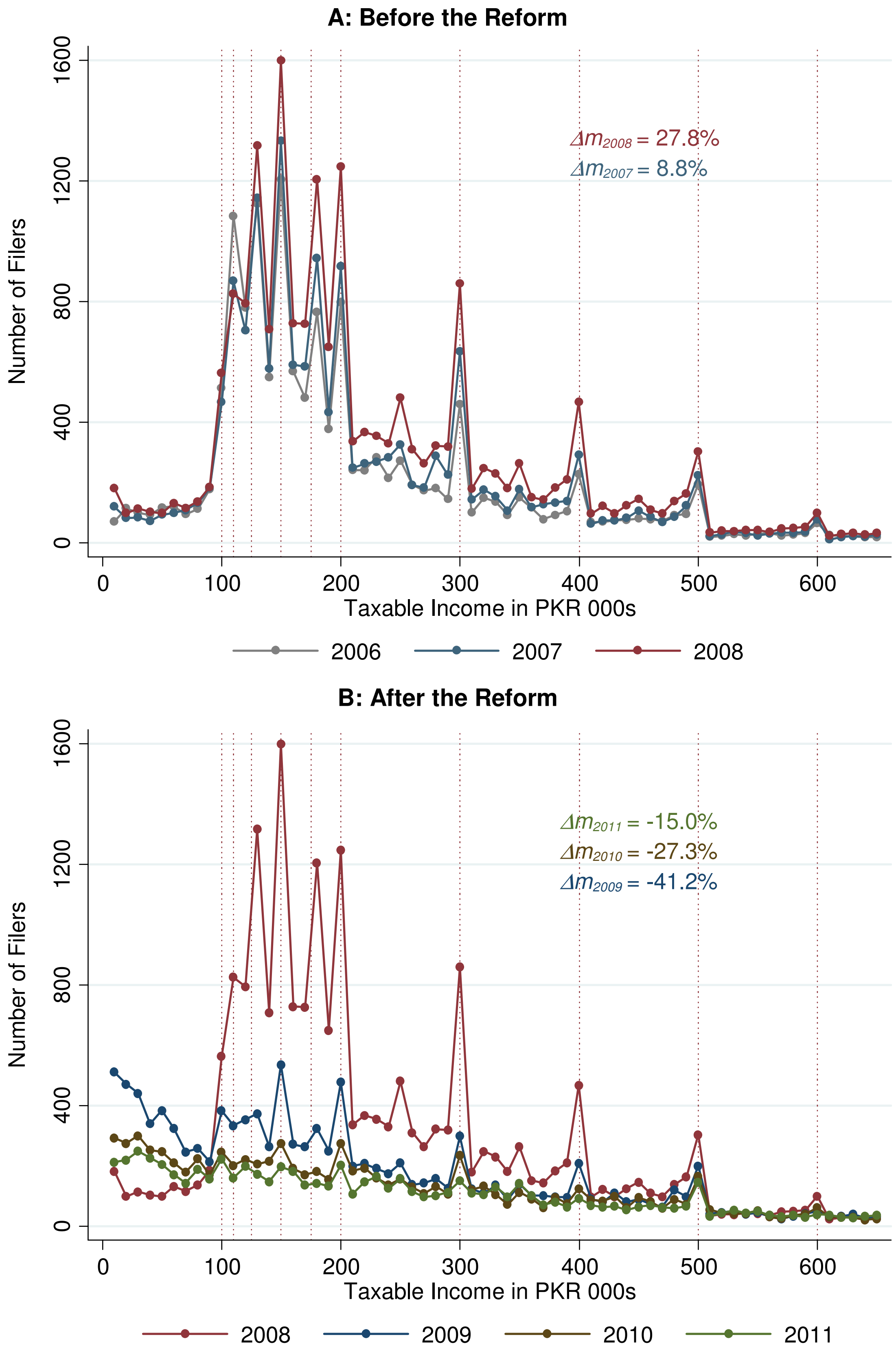
Size and Dynamics of Partnership Tax Penalty



Notes: the figure shows year-wise distribution of partnership tax penalty as a percentage of taxable income. The variable represents the additional tax that individual i experiences as a percentage of her taxable income on reporting partnership earnings as compared to if same earnings were to be reported as sole proprietorship income. The density distribution is shown in bins of size 0.83, where each bin includes the upper bound of the interval. Red vertical line demarcates the boundary below which taxpayers experience tax subsidy rather than penalty.

Figure 1.3

Taxable Income Distribution of Partnership Firms



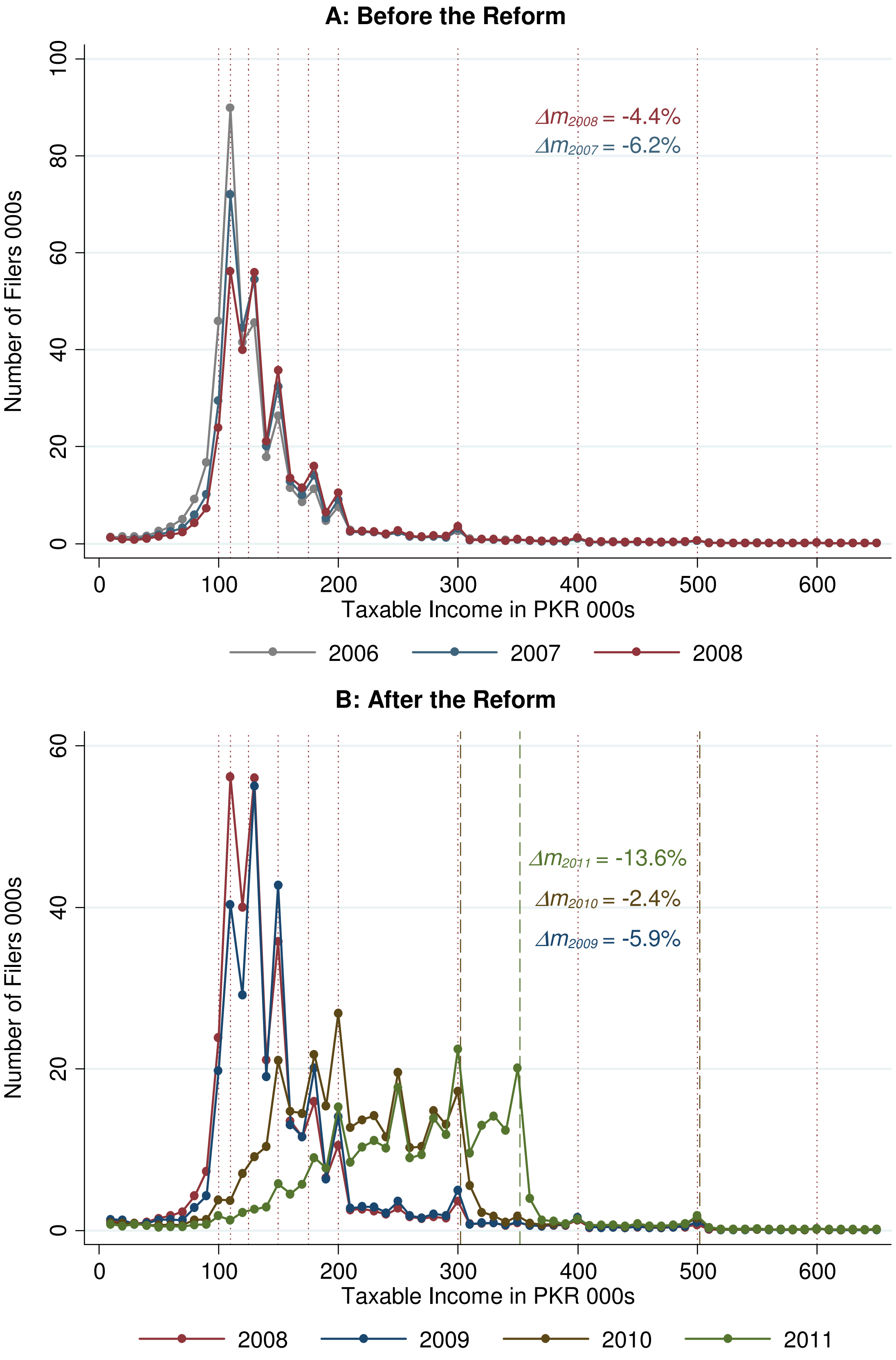
A: Before the Reform

B: After the Reform

Notes: the figure shows observed taxable income distributions of partnership firms in 2006-11. Each dot in the figure represents the upper bound of a 10,000 Rupees bin and shows the number of firms located within that bin. Notches in the 2006-08 schedule are shown by vertical dotted red lines. In Panel B, 2008 distribution is plotted again for comparison purposes. Yearly change in number of filers is represented by Δm_t, which shows the change from year *t* to *t+1* as a percentage of number of filers in year *t*.

Figure 1.4

Taxable Income Distribution of Sole Proprietorship Firms



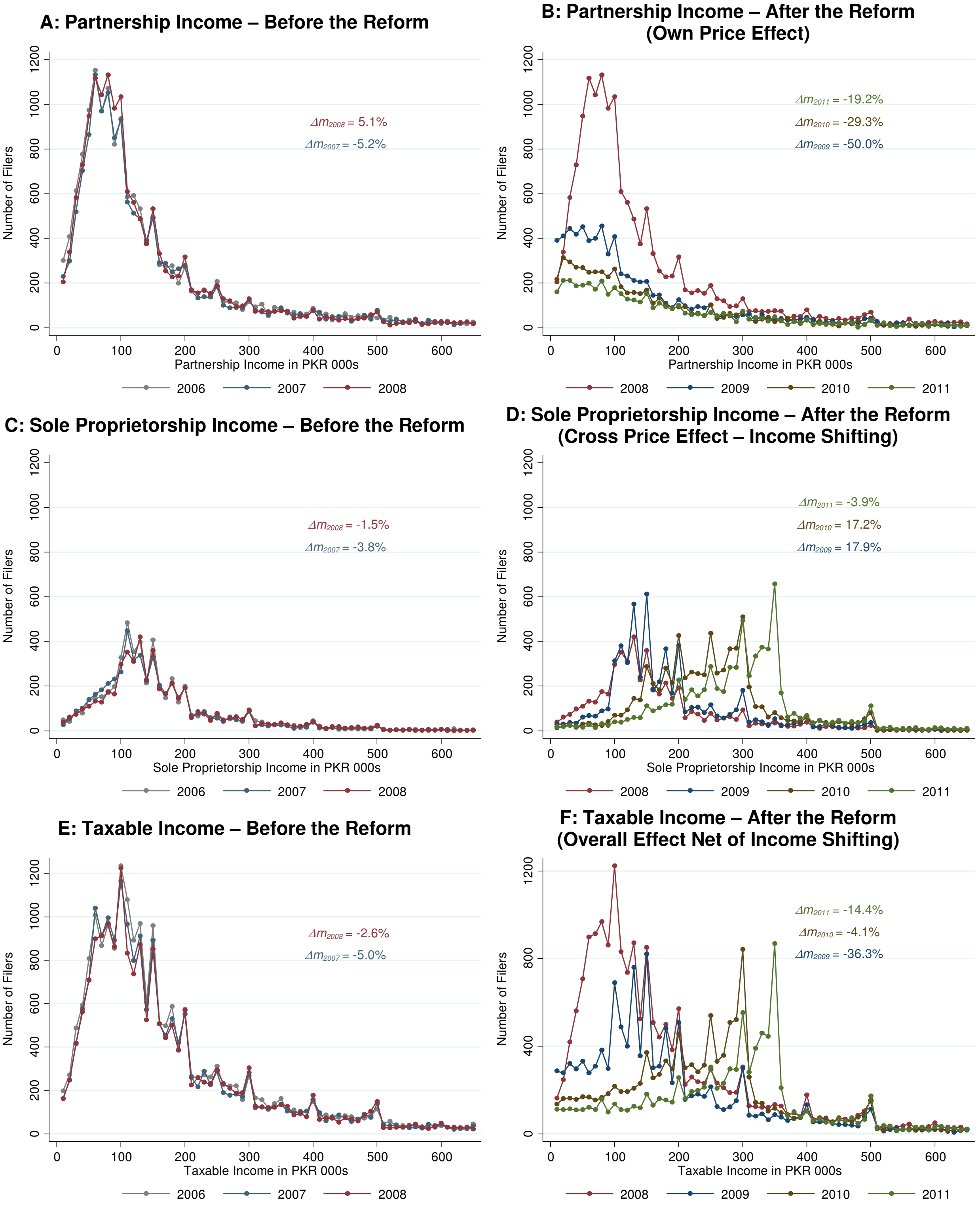
A: Before the Reform

B: After the Reform

Notes: the figure shows observed taxable income distributions of sole proprietorship firms in 2006-11. Each dot in the figure represents the upper bound of a 10,000 Rupees bin and shows the number of tax filers located within that bin. Notches in the 2006-08 schedule are shown by vertical dotted red lines. Dashed vertical lines at PKR 300,000 and 350,000 demarcate respectively the exemption threshold in 2010 and 2011; 500,000 is a notch in post-reform schedule in 2010-11. In Panel B, 2008 distribution is plotted again for comparison purposes. Yearly change in number of filers is represented by Δm_t, which shows the change from year t to $t+1$ as a percentage of number of filers in year t.

Figure 1.5

Income Shifting



A: Partnership Income – Before the Reform

B: Partnership Income – After the Reform (Own Price Effect)

C: Sole Proprietorship Income – Before the Reform

D: Sole Proprietorship Income – After the Reform (Cross Price Effect – Income Shifting)

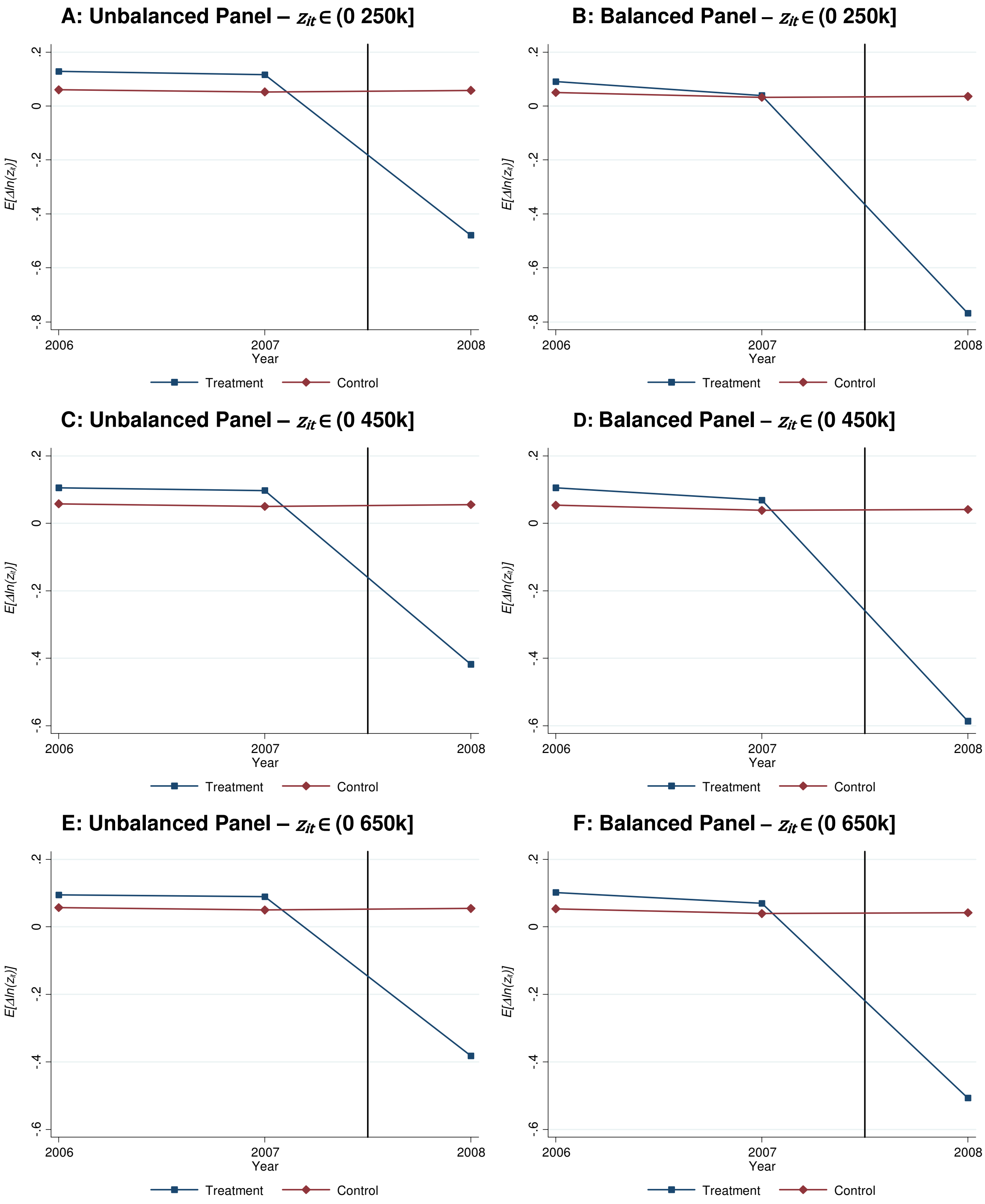
E: Taxable Income – Before the Reform

F: Taxable Income – After the Reform (Overall Effect Net of Income Shifting)

Notes: the figure shows empirical density distributions of partnership income, sole-proprietorship income and taxable income reported by individuals in their personal income tax returns in 2006-11. Taxable income for these individuals is equivalent to sum of the partnership and sole proprietorship earnings and hence its response captures overall effects of the reform net of income shifting. Each dot in the figure represents the upper bound of a 10,000 Rupees bin and shows the number of tax filers located within that bin. Left and right panels show respectively the pre-reform and post-reform distributions. Yearly change in number of filers is represented by Δm_t, which shows the change from year t to $t+1$ as a percentage of number of filers in year t.

Figure 1.6

Effects of the Tax Reform on Partnership Firms – Intensive Margin



A: Unbalanced Panel – $z_{it} \in$ (0 250k]

B: Balanced Panel – $z_{it} \in$ (0 250k]

C: Unbalanced Panel – $z_{it} \in$ (0 450k]

D: Balanced Panel – $z_{it} \in$ (0 450k]

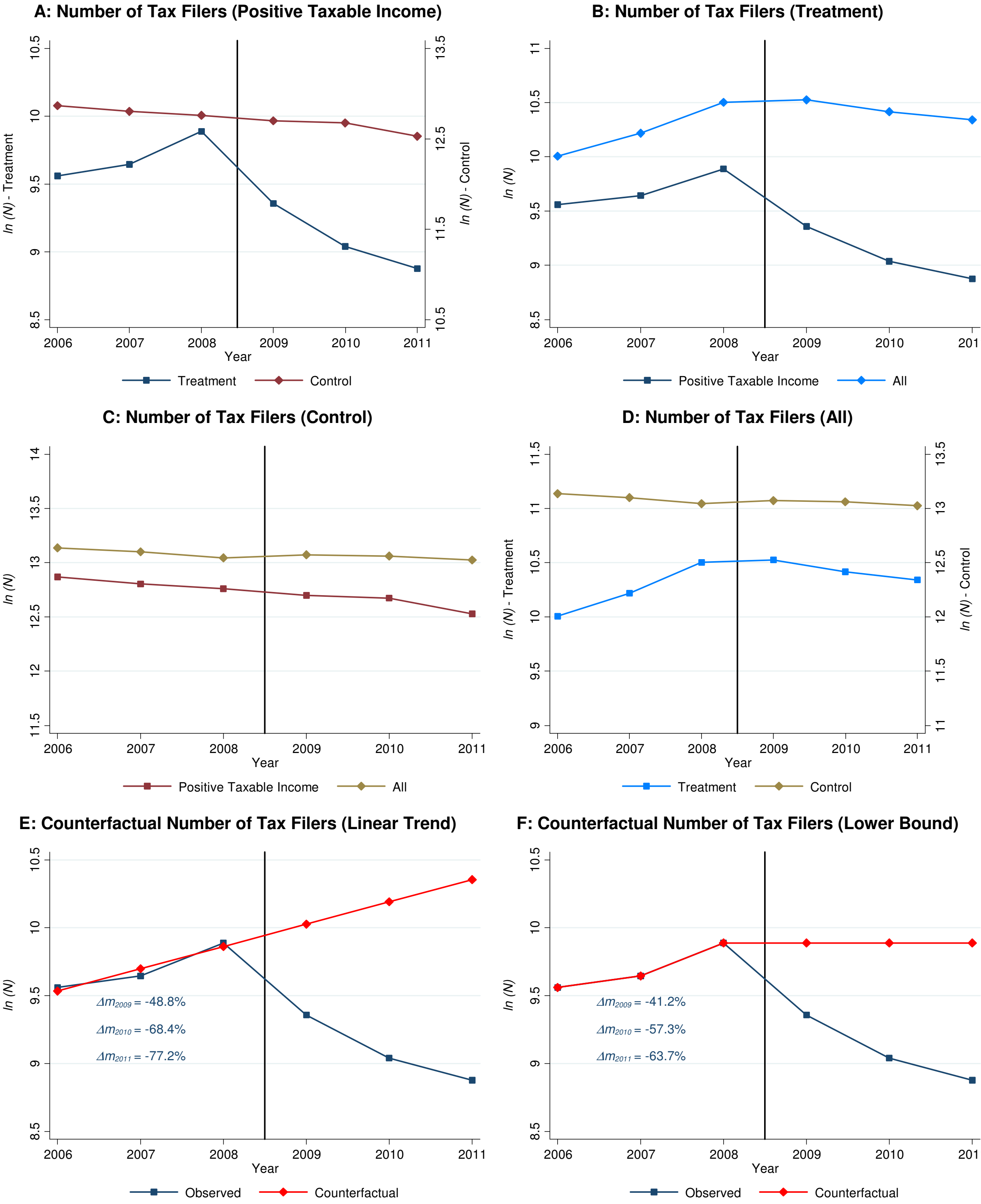
E: Unbalanced Panel – $z_{it} \in$ (0 650k]

F: Balanced Panel – $z_{it} \in$ (0 650k]

Notes: the figure shows the evolution of reported taxable income for the treatment and control groups over the years 2006-09. Treatment group in each panel consists of all partnership firms that file for tax and report earnings in the range indicated on each panel, while control group comprises all sole proprietorship firms that report taxable income in the corresponding range. Each point in the plots denotes log change in reported earnings from year *t* to *t+1* for firm *i* averaged across all firms in year *t*. Left panels include for year *t* the firms that report in two consecutive years *t* and *t+1*, while right panels include the firms that report positive taxable income in all four years in the sample. Black vertical line in each panel indicates the time from which the tax changes affect reporting behavior of the treated firms.

Figure 1.7

Effects of the Tax Reform on Partnership Firms – Extensive Margin



A: Number of Tax Filers (Positive Taxable Income)

B: Number of Tax Filers (Treatment)

C: Number of Tax Filers (Control)

D: Number of Tax Filers (All)

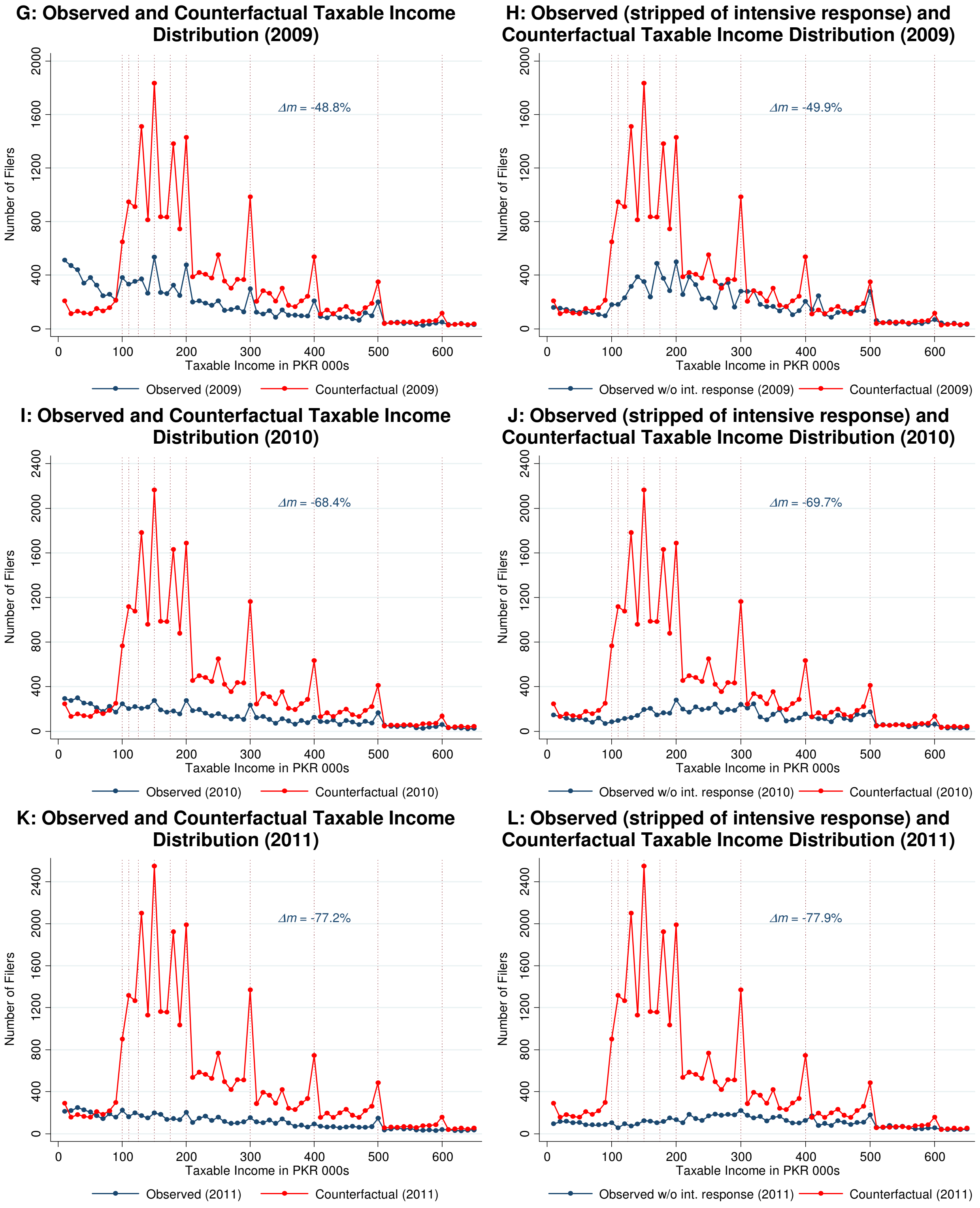
E: Counterfactual Number of Tax Filers (Linear Trend)

F: Counterfactual Number of Tax Filers (Lower Bound)

Notes: the figure shows visually the strategy to estimate the extensive margin response of partnership firms. Treatment and control groups for the figure consist respectively of partnership and sole proprietorship firms. Panel A illustrates evolution of filing for the treatment and control groups in 2006-11. Panel B plots two filing series for the treatment group: dark blue curve illustrates number of partnership firms that report positive taxable income (positive filers), while light blue curve depicts all partnership firms that file in year *t*, including those that report zero taxable earnings (all filers). Panel C plots the corresponding two series for the control group. Panel D plots all filers series for the treatment and control group together for comparison. Panel E and F show positive filers series for the treatment group along with the counterfactuals obtained from two alternative approaches. For Panel E, counterfactual is obtained by running a DD regression on two series in Panel A with separate time trends for the treatment and control groups. For Panel F, number of filers in 2008 is assumed to be counterfactual for post-reform years. Difference between counterfactual and observed number of filers for year *t* as a percentage of counterfactual number of filers for the corresponding year are indicated with Δm_t.

Figure 1.7 (Contd.)

Effects of the Tax Reform on Partnership Firms – Extensive Margin



G: Observed and Counterfactual Taxable Income Distribution (2009)

H: Observed (stripped of intensive response) and Counterfactual Taxable Income Distribution (2009)

I: Observed and Counterfactual Taxable Income Distribution (2010)

J: Observed (stripped of intensive response) and Counterfactual Taxable Income Distribution (2010)

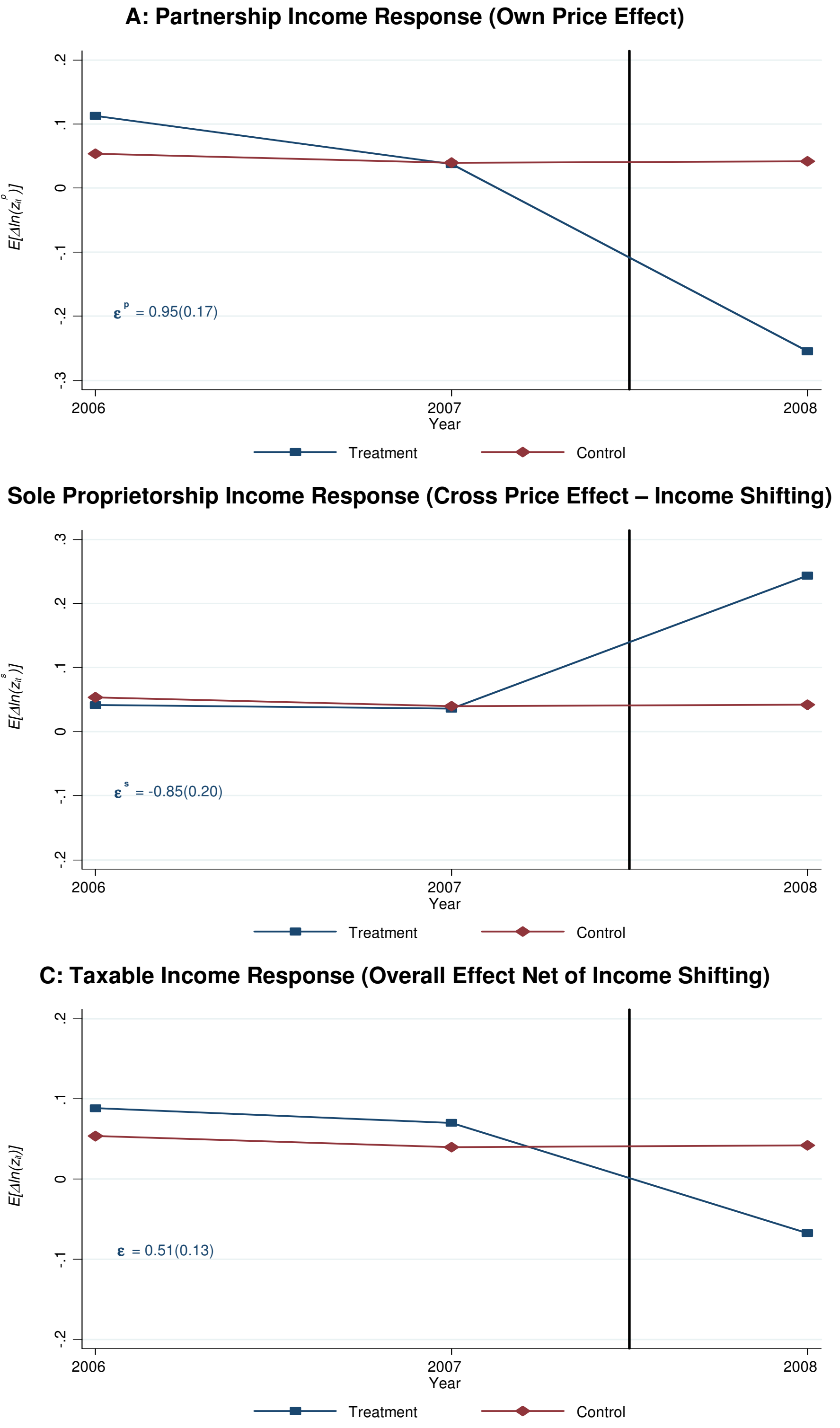
K: Observed and Counterfactual Taxable Income Distribution (2011)

L: Observed (stripped of intensive response) and Counterfactual Taxable Income Distribution (2011)

Notes: the figure shows visually the strategy to estimate extensive margin response of partnership firms. Left panels illustrate observed and counterfactual taxable income distributions of partnership firms in year *t.* Counterfactual distributions for these panels are obtained by shifting the 2008 distribution upwards proportionally to have the same mass as predicted by the DD counterfactual in Panel E of the figure. Right panels of the figure compare the observed and counterfactual distributions of treated firms but here the observed distributions have been stripped of intensive responses. The difference in number of filers between the two distributions as a percentage of number of filers in the counterfactual is denoted by *Δm*.

Figure 1.8

Effects of the Tax Reform on Partners – Intensive Margin



A: Partnership Income Response (Own Price Effect)

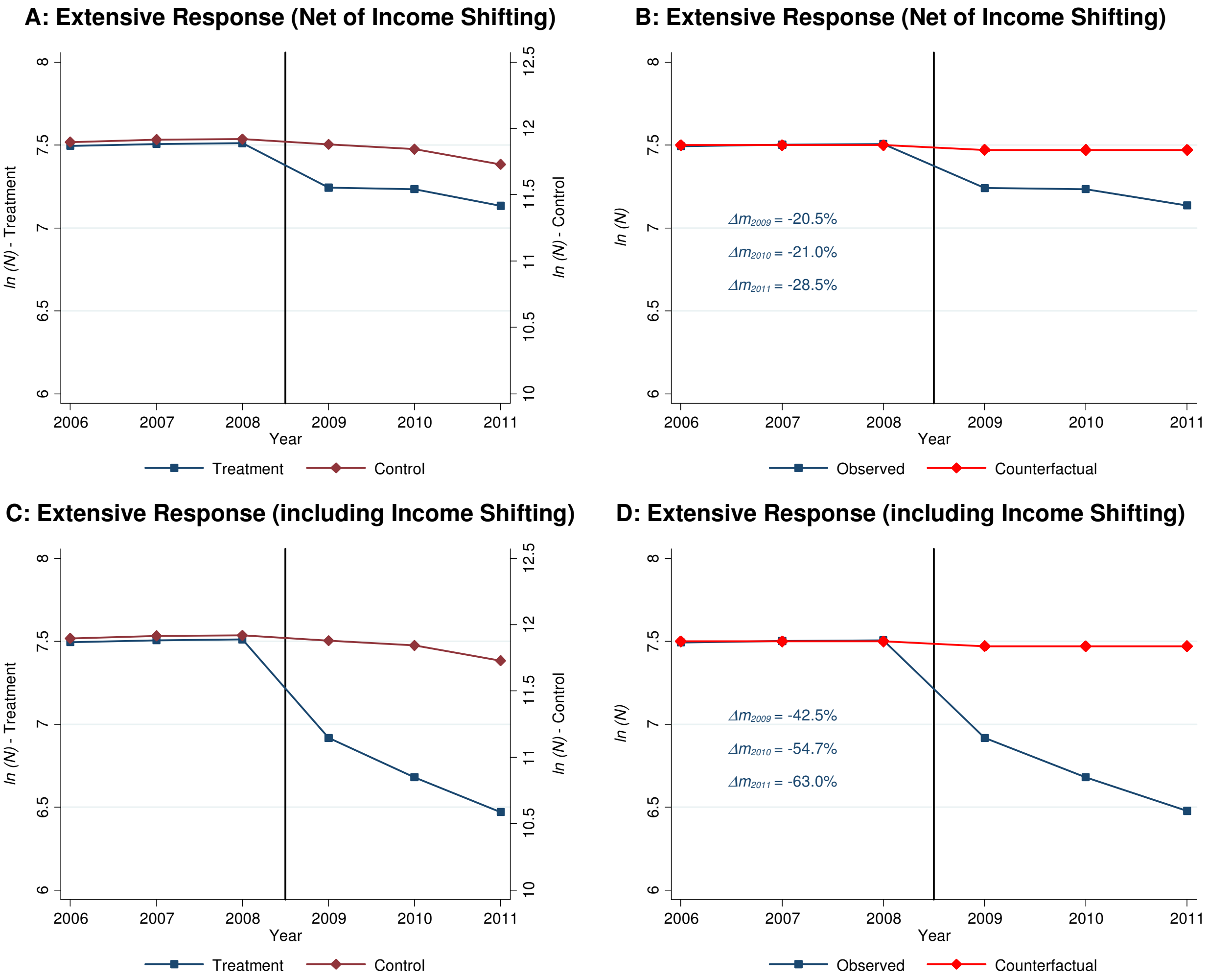
B: Sole Proprietorship Income Response (Cross Price Effect – Income Shifting)

C: Taxable Income Response (Overall Effect Net of Income Shifting)

Notes: the figure shows in Panels A, B and C respectively the evolution of partnership income, sole-proprietorship income and taxable income reported by individuals in their personal income tax returns in 2006-09. The figure is based on a balanced panel of taxpayers who file for all four years with reported taxable earnings in the range (0 650,000]. Treatment group consists of the individuals who report positive partnership income in all three years prior to the reform, while control group comprises the individuals who report zero partnership income in all years in the sample. Each point in the figure represents log change in reported income from year *t* to *t+1* for individual *i* averaged over all such individuals in year *t*. Black vertical line in each panel indicates the time from which the tax changes affect reporting behavior of the treated individuals. Elasticities given in the figure are from a 2SLS DD regressions where log change in net-of-tax rate has been instrumented with the dummy for belonging to post-reform, treatment group. Standard errors from the regression are shown in parenthesis, which are clustered at the individual level.

Figure 1.9

Effects of the Tax Reform on Partners – Extensive Margin



A: Extensive Response (Net of Income Shifting)

B: Extensive Response (Net of Income Shifting)

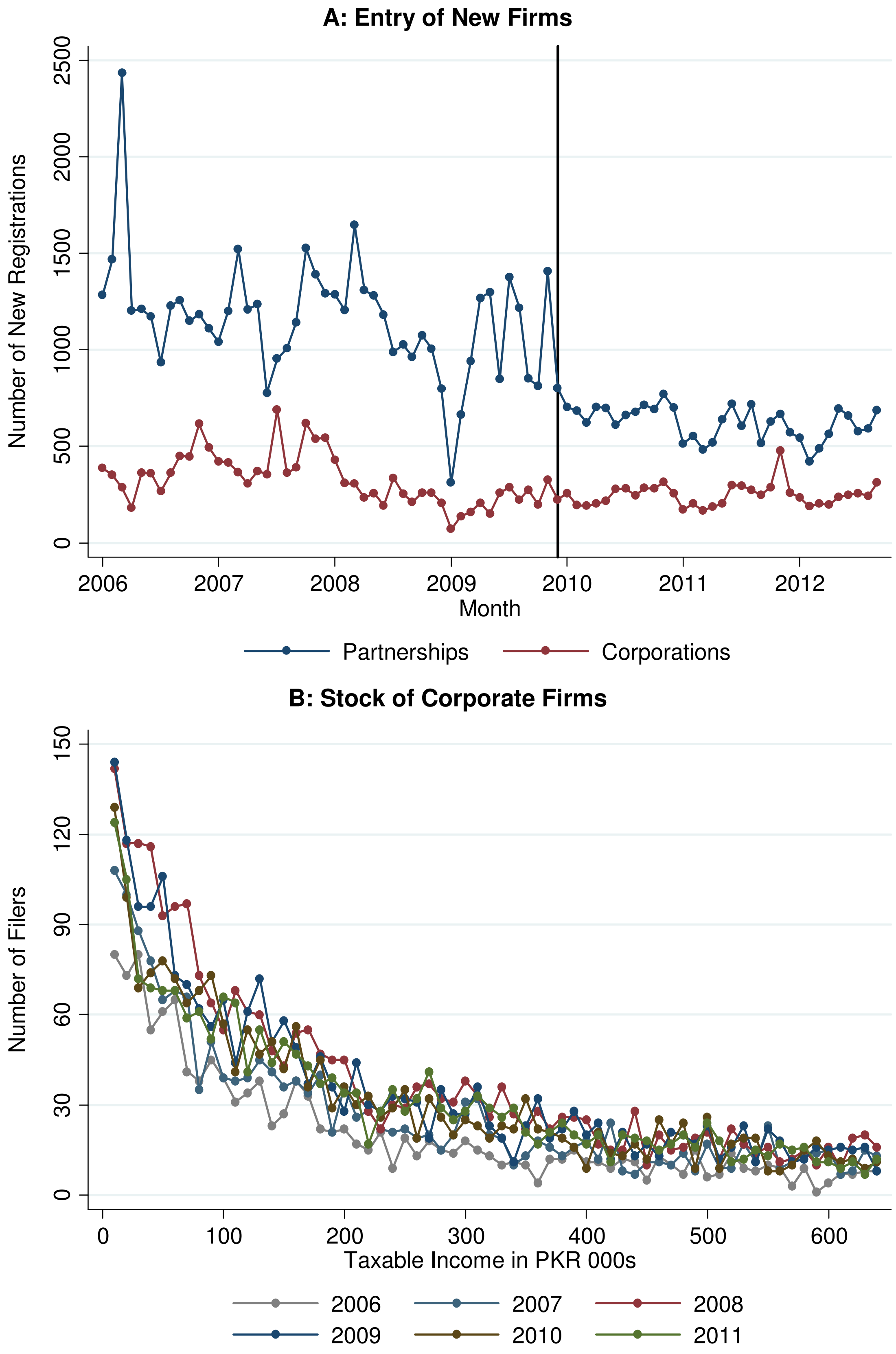
C: Extensive Response (including Income Shifting)

D: Extensive Response (including Income Shifting)

Notes: the figure shows the extensive margin response of the owners of partnership firms. The figure is based on a balanced panel of taxpayers who report in all six years 2006-11. Treatment group consists of the individuals who report positive partnership income in all three years prior to the reform, while control group comprises the individuals who report zero partnership income in all years in the sample. Panel A reports the log number of tax filers in the treatment and control groups who report positive taxable income in years 2006-11. Panel B shows the log number of tax filers in the treatment group along with a counterfactual obtained from DD regression on the two series in Panel A. Difference between the counterfactual and observed number of filers in year *t* as a percentage of the counterfactual number of filers for the corresponding year are indicated with Δm_t. Panels C and D repeat the analysis in Panels A and B, but for the treatment group consider only those individuals as *filers* who report positive partnership income in year *t*.

Figure 1.10

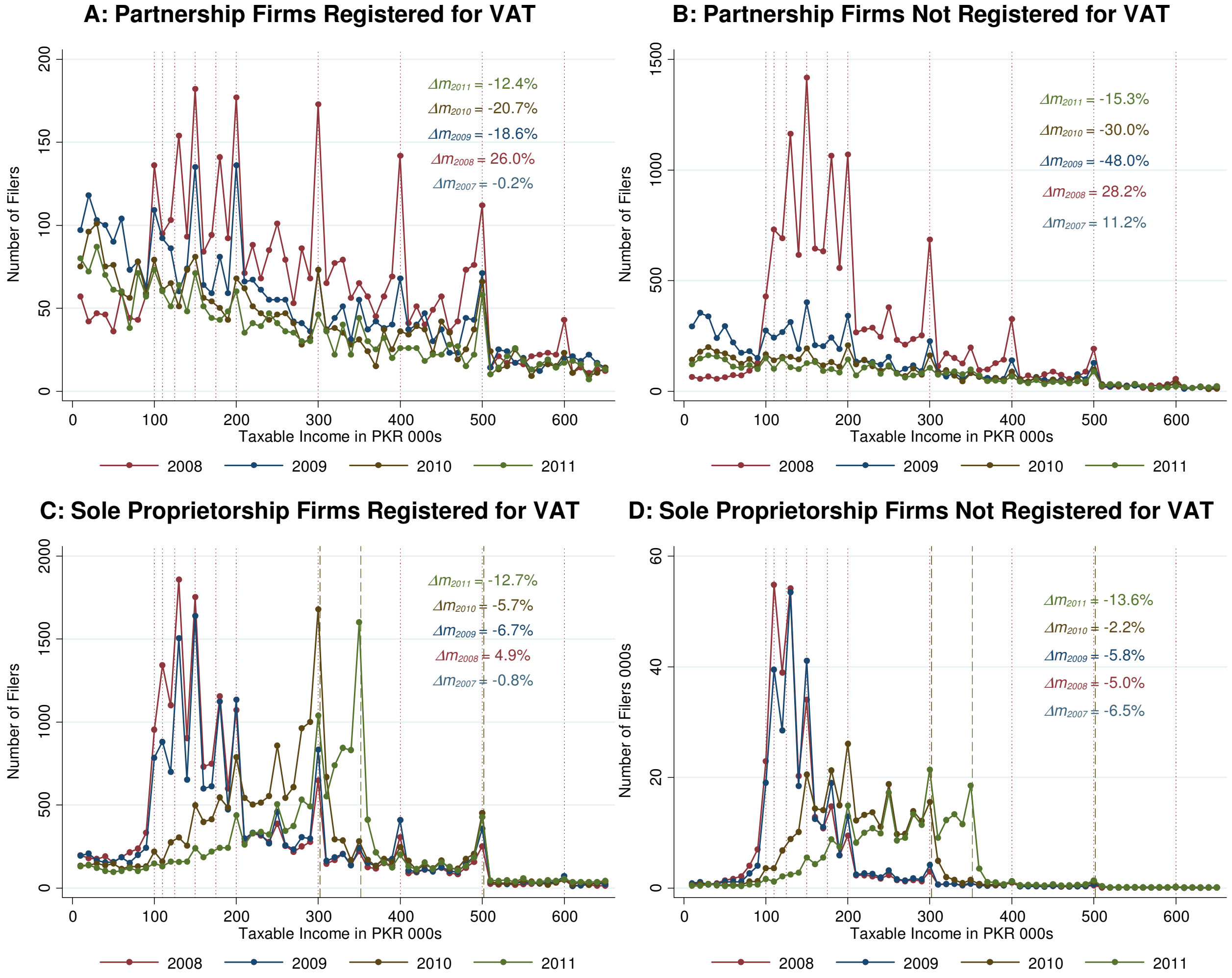
Income Shifting to Corporate Tax Base



Notes: the figure explores income shifting to corporate tax base. Panel A compares *entry* of new partnership and corporate firms. Each dot in the figure represents the number of new firms that register with the tax department in a given calendar month; year *t* in the figure denotes month July of the corresponding year. Panel B of the figure depicts the dynamics of *stock* of corporate firms. Six series in the panel plot taxable income distributions of corporate firms in 2006-11. Each dot in the plot represents the upper bound of a 10,000 Rupee bin and shows the number of corporations located within that bin.

Figure 1.11

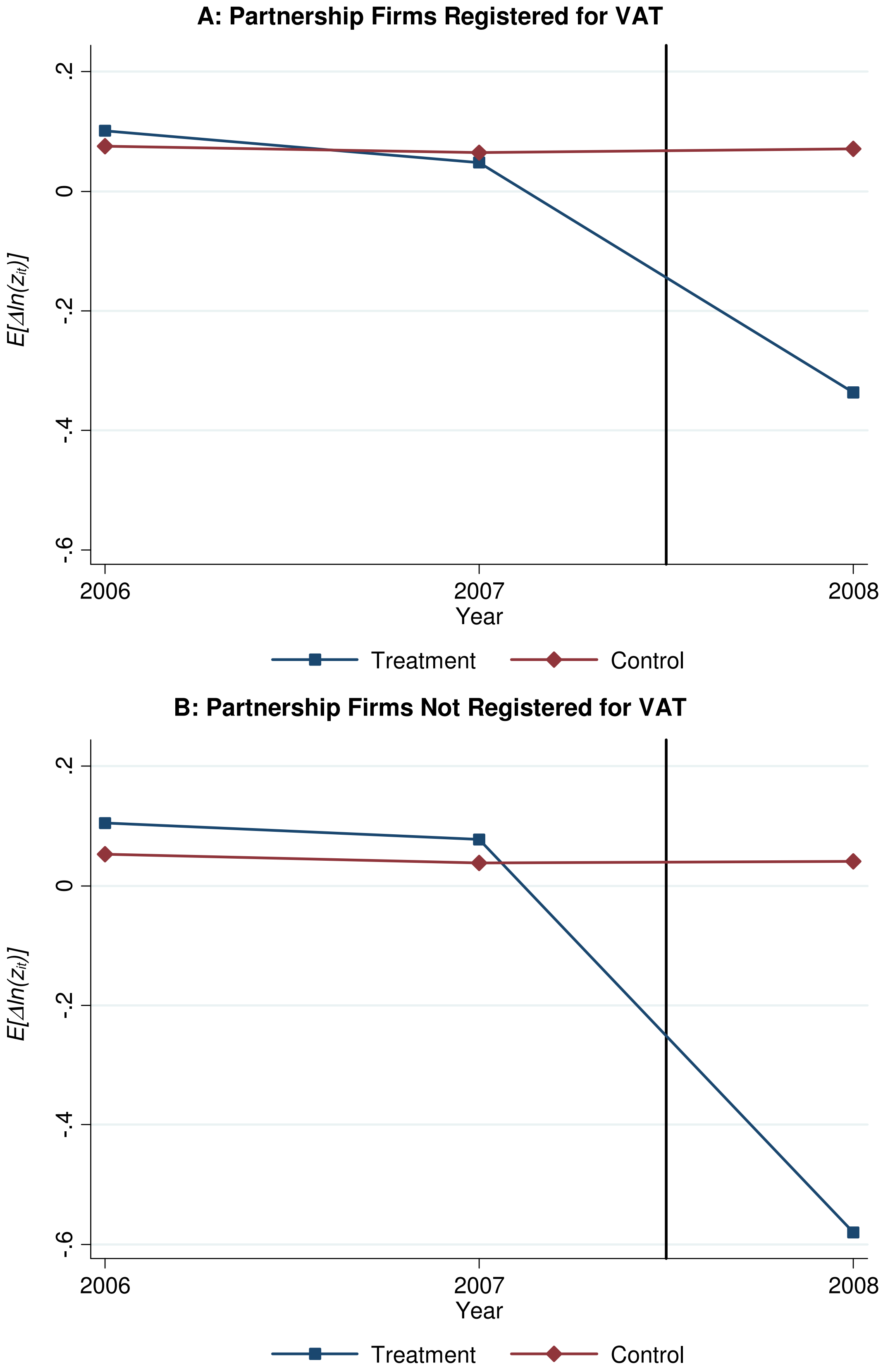
Spillover Effects on VAT Base



Notes: the figure illustrates the spillover effects of the reform on VAT base. Left panels of the figure show taxable income distributions of firms registered for VAT, while right panels depict corresponding distributions of firms not registered for VAT. Each dot in the figure represents the upper bound of a 10,000 Rupees bin and shows the number of firms located within that bin. Notches in the 2006-08 schedule are shown by vertical dotted red lines. Dashed vertical lines at PKR 300,000 and 350,000 demarcate respectively the exemption threshold in 2010 and 2011; 500,000 is a notch in post-reform schedule in 2010-11. Yearly change in number of filers is represented by Δm_t, which shows the change from year t to $t+1$ as a percentage of number of filers in year t. Empirical distribution in 2006-07 are not shown, but yearly change in number of filers for these years are included.

Figure 1.12

Do VAT Registered Firms Respond Less? Intensive Margin



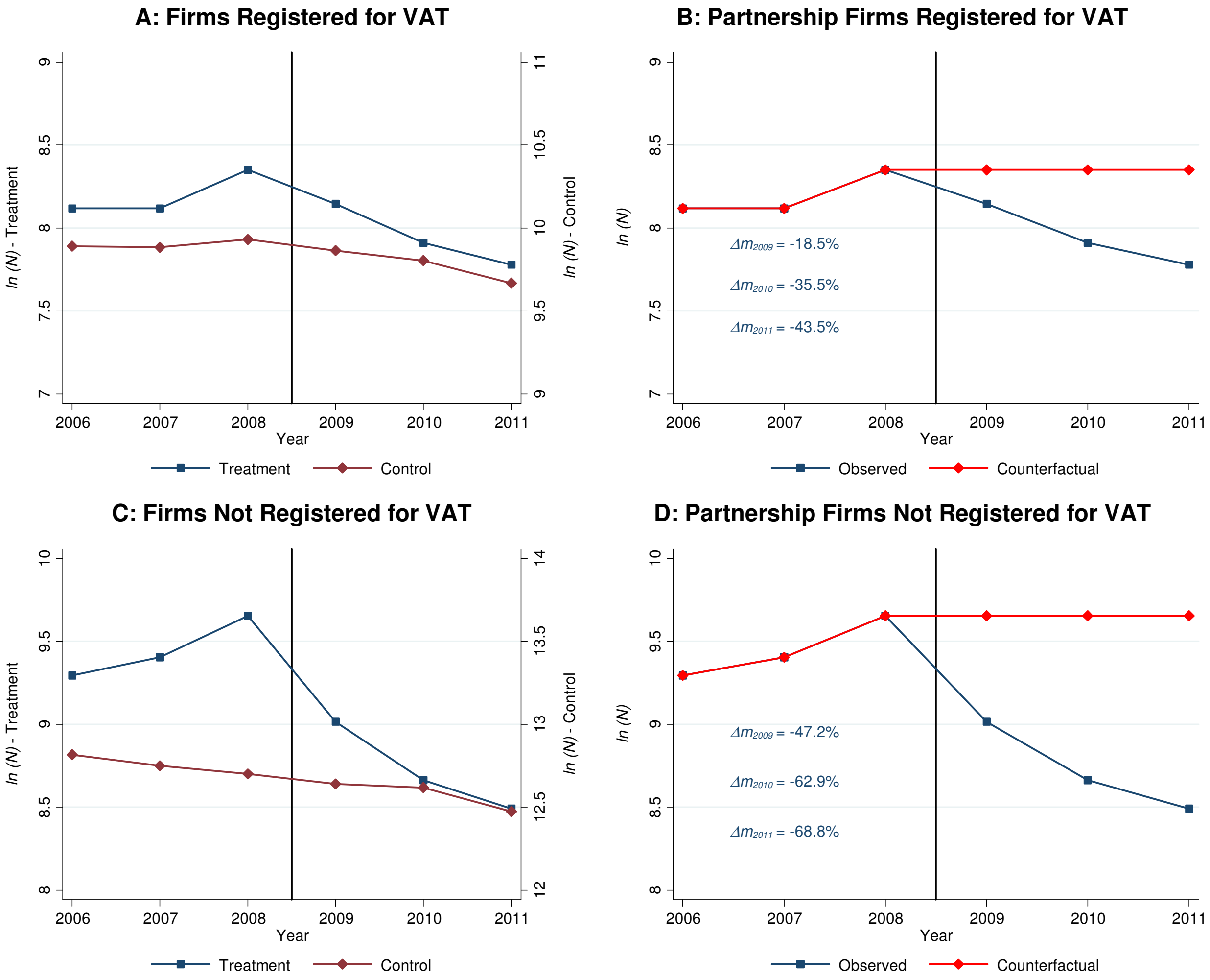
A: Partnership Firms Registered for VAT

B: Partnership Firms Not Registered for VAT

Notes: the figure investigates if VAT-registered firms respond less on the intensive margin. The figure is based on a balanced panel of firms that file in all four years (2006-09) with taxable earnings in the range (0 650k]. Treatment group in Panels A and B consist respectively of VAT-registered and VAT-unregistered partnership firms, while control group comprises the corresponding sole proprietorship firms. Each point in the figure represents log change in reported income from year *t* to *t+1* for firm *i* averaged over all filers in year *t*. Black vertical line in each panel indicates the time from which the tax changes affect reporting behavior of the treated firms.

Figure 1.13

Do VAT Registered Firms Respond Less? Extensive Margin



A: Firms Registered for VAT

B: Partnership Firms Registered for VAT

C: Firms Not Registered for VAT

D: Partnership Firms Not Registered for VAT

Δm_{2009} = -18.5%

Δm_{2010} = -35.5%

Δm_{2011} = -43.5%

Δm_{2009} = -47.2%

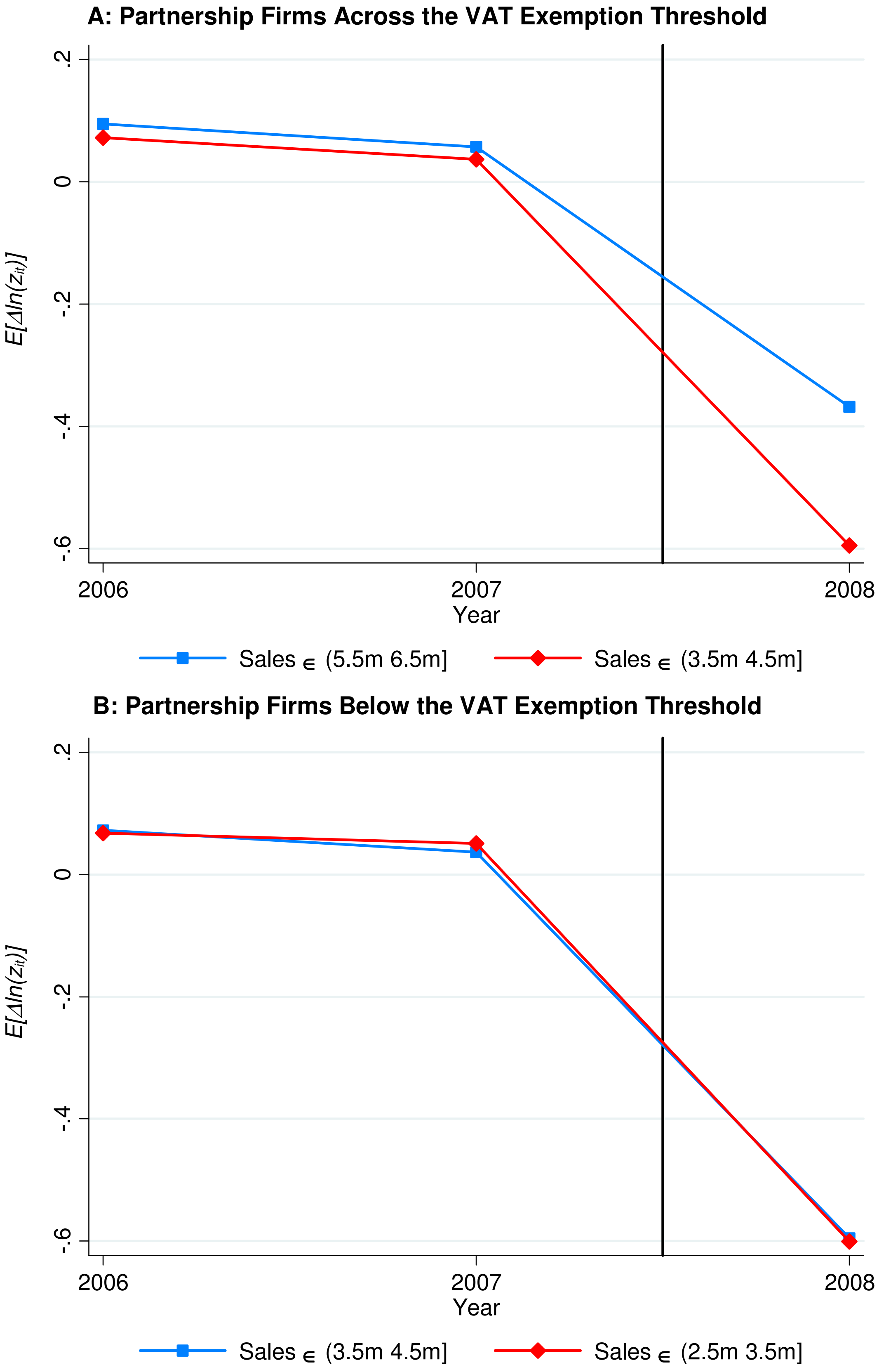
Δm_{2010} = -62.9%

Δm_{2011} = -68.8%

Notes: the figure compares extensive margin responses of partnership firms registered for VAT with partnership firms not registered for VAT. Treatment and control groups for the figure consist respectively of partnership and sole proprietorship firms. Panel A reports log number of firms in the treatment and control groups that are registered to remit VAT on their sales and report positive taxable income in 2006-11. Panel B shows such firms in the treatment group along with a counterfactual that assumes that absent the tax changes number of such firms would have stayed at the pre-reform level. Difference between counterfactual and observed number of filers for year *t* as a percentage of counterfactual number of filers for the corresponding year are indicated with Δm_t. Panels C and D repeat the analysis in Panels A and B for firms not registered for VAT.

Figure 1.14

Is VAT Causal?



Notes: the figure explores if registering for VAT makes the firms more tax compliant. Panel A of the figure compares short-term taxable income responses – which arguably capture tax evasion – of firms on both sides of the VAT exemption threshold (annuals sales not more than Rs. 5 million). Panel B makes similar comparison for firms with sales below the exemption. Treatment and control groups in each panel consist respectively of the partnership and sole proprietorship firms that file for tax and report earnings in the range (0 650k]. Each point in the panels represents log change in reported income from year t to $t+1$ for firm i averaged over all filers in year t. Black vertical line in each panel indicates the time from which the tax changes affect reporting behavior of the treated firms.

Table 1.1

Size and Dynamics of Partnership Tax Penalty

Year	**Partnership Tax Penalty (PKR)**	**Partnership Tax Penalty (Percentage of Taxable Income)**
(1)	(2)	(3)
2006	7,886	4.1
	(15,118)	(4.7)
2007	6,747	3.7
	(13,149)	(4.1)
2008	6,074	3.7
	(12,228)	(4.0)
2009	22,246	16.5
	(18,030)	(9.3)
2010	28,495	15.4
	(24,457)	(9.8)
2011	31,467	14.8
	(26,675)	(10.1)

Notes: the table presents the estimates of size and dynamics of partnership tax penalty in 2006-11. Column (2) of the table shows the additional tax that individual *i* experiences on reporting partnership earnings as compared to if same earnings were to be reported as sole proprietorship income (partnership tax penalty), averaged over all individuals who report positive partnership earnings. Column (3) shows partnership tax penalty as a percentage of taxable income of individual *i*, averaged over all individuals who report positive partnership earnings. Standard errors are in parenthesis.

Table 1.2

Intensive Margin Elasticities for Partnership Firms

Taxable income (≤)	**Unbalanced Panel**			**Balanced Panel**		
	# Obs		**ε**	**# Obs**		**ε**
	Control	**Treatment**		**Control**	**Treatment**	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
250,000	779,124	19,438	**2.306** (0.089)	484,374	3,915	**2.939** (0.157)
350,000	806,825	24,260	**2.246** (0.077)	514,740	6,054	**2.671** (0.129)
450,000	816,231	27,068	**2.169** (0.073)	525,264	7,503	**2.604** (0.121)
550,000	820,477	28,798	**2.115** (0.073)	530,631	8,670	**2.495** (0.119)
650,000	821,916	29,538	**2.073** (0.073)	531,975	9,207	**2.402** (0.121)

Notes: the table presents intensive margin elasticity estimates from 2SLS regressions. Sample includes the partnership (treatment) and sole proprietorship firms (control) in Pakistan that report taxable earnings in the interval indicated in column (1) in 2006-09. Column (4) and (7) report the coefficients on Δlog net-of-tax rate in differences-in-differences regressions, where Δlog net-of-tax rate has been instrumented in the first stage with a dummy for belonging to post-reform, treatment group. Sample for columns (2) to (4) includes the firms that report for two consecutive years *t* and *t+1* and for columns (5) to (7) only the firms that report for all four years in the sample. Standard errors are in parenthesis, which are clustered at the level of firm. All coefficients are significant at 1% level.

Table 1.3

Extensive Margin Elasticities for Partnership Firms

	2009			**2010**			**2011**		
Taxable Income (000s)	**Observed Distribution Stripped of Intensive Response (#Obs)**	**Counterfactual Distribution (#Obs)**	***η***	**Observed Distribution Stripped of Intensive Response (#Obs)**	**Counterfactual Distribution (#Obs)**	***η***	**Observed Distribution Stripped of Intensive Response (#Obs)**	**Counterfactual Distribution (#Obs)**	***η***
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A: Linear Trend as Counterfactual									
0 - 200	4,679	13,221	**2.438**	2,678	15,586	**3.125**	2,016	18,374	**3.360**
			(0.222)			(0.260)			(0.314)
200 - 300	2,688	4,518	**1.716**	2,004	5,326	**2.643**	1,664	6,279	**3.114**
			(0.279)			(0.238)			(0.263)
300 - 400	1,810	2,581	**1.486**	1,482	3,043	**2.552**	1,372	3,587	**3.072**
			(0.341)			(0.299)			(0.289)
400 - 500	1,497	1,597	0.407	1,249	1,883	**2.186**	1,129	2,220	**3.191**
			(0.300)			(0.330)			(0.360)
B: Pre-reform Level as Counterfactual (Lower Bound)									
0 - 200	4,679	11,521	**2.241**	2,678	11,521	**2.896**	2,016	11,521	**3.113**
			(0.220)			(0.258)			(0.310)
200 - 300	2,688	3,937	**1.344**	2,004	3,937	**2.080**	1,664	3,937	**2.446**
			(0.279)			(0.238)			(0.263)
300 - 400	1,810	2,249	**0.971**	1,482	2,249	**1.697**	1,372	2,249	**1.940**
			(0.322)			(0.299)			(0.289)
400 - 500	1,497	1,392	-0.49	1,249	1,392	**0.667**	1,129	1,392	**1.227**
			(0.300)			(0.330)			(0.360)

Notes: the table presents extensive margin elasticity estimates for partnership firms. Column (3), (6) and (9) of the table show the number of tax filers in the counterfactual distribution - the distribution that would have observed had there been no tax changes. Columns (2), (5) and (7) of the table report number of filers in the observed distribution that has been stripped of intensive responses. This distribution would have been observed had there been no response to the tax rate changes on the intensive margin. Elasticity estimates in columns (4), (7) and (10) are from a simple regression of log difference in number of filers in each bin of the two distributions against log changes in net-of-tax rate experienced by the tax filers in that bin. Standard errors are in parenthesis. Estimates in Panels A and B are based on the two alternative approaches to estimate the counterfactual. Coefficients significant at 5% level are shown in bold.

Table 1.4

Intensive Margin Elasticities for Partnership Firms by VAT Registration

Taxable income (≤)	**Firms Not Registered for VAT**			**Firms Registered for VAT**		
	# Obs		**ε**	**# Obs**		**ε**
	Control	**Treatment**		**Control**	**Treatment**	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
250,000	468,144	3,168	**3.191** (0.169)	16,821	771	**1.922** (0.381)
350,000	493,236	4,638	**2.945** (0.144)	22,134	1,425	**1.756** (0.274)
450,000	501,009	5,490	**2.885** (0.136)	24,945	2,019	**1.764** (0.244)
550,000	504,501	6,180	**2.771** (0.132)	26,868	2,493	**1.723** (0.225)
650,000	505,317	6,414	**2.747** (0.133)	27,393	2,799	**1.502** (0.232)

Notes: the table presents intensive margin elasticity estimates from 2SLS regressions. Sample has been stratified by VAT-registration, and includes the partnership (treatment) and sole proprietorship (control) firms that file for tax in all four years 2006-09 and report taxable earnings in the interval indicated in column (1). Column (4) and (7) report the coefficients on log change in net-of-tax rate in differences-in-differences regressions, where log change in net-of-tax rate has been instrumented in the first stage with a dummy for belonging to post-reform treatment group. Standard errors are in parenthesis, which are clustered at the level of firm. All coefficients are significant at 1% level.

Table 1.5

Which Firms Respond to the Reform? Intensive Margin

	Dependent Variable: Log Change in Reported Earnings							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Partnership x After	-0.5400***	-0.6190***	-0.5620***	-0.4425***	-0.6547***	-0.6140***	-0.5767***	-0.6357***
	(0.0252)	(0.0326)	(0.0274)	(0.0404)	(0.0273)	(0.0278)	(0.0260)	(0.0529)
Partnership x After x Large		0.3467***						0.2043***
		(0.0396)						(0.0510)
Partnership x After x Age			0.0892**					0.0833**
			(0.0360)					(0.0419)
Partnership x After x Electronic Return Filer				-0.1207***				-0.2259***
				(0.0401)				(0.0514)
Partnership x After x Third Party Reporting					0.4896***			0.4561***
					(0.0307)			(0.0397)
Partnership x After x Registered for VAT						0.2436***		0.1580***
						(0.0330)		(0.0388)
Partnership x After x Withholding Agent							0.3039***	0.0352
							(0.0413)	(0.0581)
R-squared	0.0273	0.0296	0.0282	0.0275	0.0311	0.0284	0.0282	0.0352
Observations	540,705	418,716	522,126	540,705	540,705	540,705	540,705	410,247

Notes: the table explores determinants of tax evasion. Columns (1) - (8) present estimates from difference-in-difference regressions, which include firm characteristics interactions. The regressions are based on a balanced panel sample that comprises the partnership (treatment) and sole proprietorship (control) firms that report taxable earnings in the interval (0 650k] in 2006-09. The double Interaction term, Partnership x After, arguably captures the tax evasion of partnership firms induced by the reform. Triple interaction terms reflect how the evasion varies with firm observables. Details of the firm characteristics variables are given in appendix. Standard errors are in parenthesis, which are clustered at the level of firm.

*, **, *** represent statistical significance at the 10%, 5%, and 1% percent levels respectively.

Table 1.6

Which Firms Respond to the Reform? Extensive Response

	2009	2010	2011	All
	(1)	(2)	(3)	(4)
Size Quartile 2	-0.0280**	-0.0266**	-0.0117	-0.0199**
	(0.0132)	(0.0117)	(0.0113)	(0.0101)
Size Quartile 3	-0.1599***	-0.1471***	-0.1158***	-0.1116***
	(0.0138)	(0.0133)	(0.0132)	(0.0121)
Size Quartile 4	-0.1608***	-0.1627***	-0.1512***	-0.1485***
	(0.0160)	(0.0163)	(0.0164)	(0.0155)
Electronic Return Filer	-0.2491***	-0.2669***	-0.2595***	-0.2076***
	(0.0169)	(0.0125)	(0.0110)	(0.0090)
Third Party Quartile 2	0.0233*	0.0160	0.0207*	0.0087
	(0.0133)	(0.0122)	(0.0119)	(0.0109)
Third Party Quartile 3	-0.0830***	-0.0716***	-0.0557***	-0.0518***
	(0.0134)	(0.0132)	(0.0129)	(0.0121)
Third Party Quartile 4	-0.1390***	-0.1164***	-0.1073***	-0.0863***
	(0.0138)	(0.0142)	(0.0143)	(0.0136)
Registered for VAT	-0.0999***	-0.1293***	-0.0886***	-0.0734***
	(0.0119)	(0.0120)	(0.0121)	(0.0115)
Withholding Agent	-0.0902***	-0.0682***	-0.0832***	-0.0611***
	(0.0131)	(0.0146)	(0.0149)	(0.0146)
Controls	YES	YES	YES	YES
Number of Firms that Exit or Break Up	4,583	6,533	7,024	7,942
R-squared	0.2548	0.2512	0.2211	0.1996
Observations	11,055	11,055	11,055	11,055

Notes: the table explores determinants of extensive response to the reform. Columns (1) - (4) report coefficients of OLS regressions of a dummy indicating firms that report positive taxable income in 2008 but do not in year t on dummy covariates and a rich set of controls comprising (i) tax office fixed effects (fifteen categories) (ii) industry fixed effects (six-digit industry code the firm belongs to) and (iii) age fixed effects (10 categories). All variables are introduced non-parametrically. The details of dummy covariates of interest for which coefficients are shown in the table are in appendix. Standard errors are shown in parentheses. Sample for the regression consists of all partnership firms that report positive taxable income in 2008 and for which information on all included variables (339 dummies) are available. Bottom row also shows number of firm that report zero taxable earnings in year t indicating the exit or break up of the firm.

*, **, *** represent statistical significance at the 10%, 5%, and 1% percent levels respectively.

Production vs Revenue Efficiency With Limited Tax Capacity: Theory and Evidence From Pakistan

2.1 Introduction

A central result in public economics, the *production efficiency theorem* (Diamond & Mirrlees 1971), states that tax systems should leave production undistorted even in second-best environments. This result permits taxes on consumption, wages and profits, but precludes taxes on intermediate inputs, turnover and trade. The theorem has been hugely 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 theorem considers 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 larger revenue efficiency. While there is some theoretical work along these lines (e.g. Emran & Stiglitz 2005; Gordon & Li 2009), there is virtually no empirical evidence on the trade-off between production and revenue efficiency in the choice of tax instruments.

To address this question empirically, we need simultaneous variation in tax instruments that vary with respect to their production efficiency properties (such as switches between two instruments). This is more challenging than the usual search for variation in tax rates for a given tax instrument, because we are interested in comparing instruments that apply to different tax bases and often to very different taxpayer populations. A few of studies have taken a macro cross-country approach focusing on trade vs domestic taxes (Baunsgaard & Keen 2010; Cage & Gadenne 2012). This paper proposes instead a micro approach that exploits 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.[28] This policy has been motivated by the idea

[28] Such minimum tax schemes have been implemented in numerous developing countries, including Argentina, Bolivia, Cambodia, Cameroon, Chad, Colombia, Democratic Republic of the Congo, Ecuador,

that the broader turnover base is harder to evade, an argument 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 for the 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 more revenue efficiency. Compared to existing bunching approaches (Saez 2010; Chetty *et al.* 2011; Kleven & Waseem 2013) a conceptual contribution is to develop a method that exploits the simultaneous discontinuity in the tax rate and the tax base.

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 first a stylized comparison between a turnover tax and a pure profit tax on an individual firm. Because turnover is a much broader base than profits, minimum tax schemes are always associated with very small turnover tax rates as compared to profit tax rates. For example, in our empirical application to Pakistan, the turnover tax rate is never above 1% while the profit tax rate is 20-35%. The low turnover tax rate implies that this tax introduces only a *small* distortion of 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.[29] Hence, the simultaneous changes in tax base and tax rate at the kink offset each other to produce a very small change in *real* incentives for the individual firm. 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. We show in the paper that this basic argument is robust to a number of generalizations, including real distortions of turnover taxes driven by cascading[30] and distortionary profit taxes levied

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.

[29] In general, both taxes may create significant distortions at the *extensive* margin (such as informality, sector and location choices), which depends on the effective *average* tax rate rather than the effective *marginal* tax rate. In particular, a small tax rate on turnover may create a large average tax rate on profits (tax liability as a share of profits) due to the broadness of the turnover base. However, since tax liability is continuous at the minimum tax kink, there will be no extensive responses to the kink and our bunching estimates are not affected by such responses.

[30] While a small turnover tax introduces only a small firm-level distortion of real production at the intensive margin, there may still be significant economy-wide distortions because of cascading—taxing

on bases that deviate from pure economic rent.[31] We develop a simple model allowing us to put bounds on the evasion response using bunching at the minimum tax kink under different assumptions about the real output elasticity. Due to the weak real incentive, the bounds on the evasion response are extremely tight under a very wide range of real output elasticities.

We use administrative data from the Federal Board of Revenue in Pakistan to analyze the responses by Pakistani firms to the minimum tax regime. The data contains all corporate tax returns between 2006 and 2010, which are predominantly filed electronically, contributing to the quality of the data in this context. Our main empirical findings are the following. First, we observe large and sharp bunching in reported profit rates around the threshold below which the turnover tax applies. We exploit variation in the minimum tax kink over time and across firms to confirm that the excess bunching is indeed a response to the tax system, including a temporary elimination of the minimum tax scheme as well as differences in the size and location of the kink for different populations of firms. These findings provide compelling non-parametric evidence that firms respond to the minimum tax incentives in the way that our theory predicts, and the presence of weak real incentives around the kink suggests that evasion is an important part of the story. To explore the role of evasion, we also show that firms with greater evasion opportunity—either smaller firms or firms with less activities subject to a paper trail—bunch more strongly at the minimum tax kink.

Second, we combine our non-parametric bunching evidence with a simple conceptual framework in order to bound the evasion response to switches between profit and turnover taxes under different assumptions on the real output elasticity. We find that turnover taxes reduce evasion by up to 60-70% of corporate income compared to profit taxes. The evasion estimates are very robust to the size of the real output elasticity, because the smallness of real incentives around the kink implies that real responses contribute very little to bunching even under very large elasticities. Third, we use our empirical estimates as sufficient statistics in an analysis of the optimal choice of tax base and tax rate in an environment with limited tax capacity. We find that welfare can be increased by moving away from a pure profit tax towards a much broader base that is closer (but not

the same item multiple times—through the production chain. Cascading effects of multiple-stage production can imply effective distortions far higher than statutory tax rates (Keen 2013). Importantly, such general equilibrium cascading effects do not generate *bunching* at the minimum tax kink (a firm will not escape such general equilibrium effects by bunching at the kink) and so will not be captured by our estimates. The absence of cascading effects in our bunching estimates is a great advantage for our ability to identify evasion.

[31] In general, actual corporate income taxes do not correspond to taxes on pure economic profits, and so may be associated with significant real distortions (e.g. Hassett & Hubbard 2002; Devereux & Sorensen 2006; Auerbach *et al.* 2010). By itself, this effect makes a profit tax *more* distortionary compared to a turnover tax. We show that this creates real production incentives around the minimum tax kink that move firms *away* from the kink, and therefore reinforces our argument that bunching is not driven by real responses.

identical) to turnover, because the loss of production efficiency is more than compensated for by the increase in tax compliance. It is in general not optimal to go all the way to a pure turnover tax in our framework, but the administrative simplicity of a pure turnover tax could further tip the balance in practice. Overall, our findings demonstrates that governments with limited tax capacity face an important trade-off between production efficiency and revenue efficiency that has first-order implications for the choice of tax instruments.

Our paper contributes to several literatures. First, we contribute to an emerging empirical literature on public finance and development using administrative microdata (e.g. Kumler *et al.* 2012; Kleven & Waseem 2013; Pomeranz 2013). Second, a theoretical literature has studied the implications of limited tax capacity for optimal taxation (Emran & Stiglitz 2005; Keen 2008; Boadway & Sato 2009; Gordon & Li 2009; Kleven *et al.* 2009; Besley & Persson 2011; Dharmapala *et al.* 2011). While most of these papers study movements between the formal and informal sectors, our paper studies corporate evasion at the intensive margin and derives simple expressions for optimal tax policy that depend on parameters which we estimate.[32]

Third, a vast literature studies the determinants of tax evasion (see Andreoni *et al.* 1998 and Slemrod & Yitzhaki 2002 for surveys). A large part of this literature has used macroeconomic indicators (money supply, aggregate electricity demand etc.), survey data on consumption and income, or audit data to estimate the extent of tax evasion (see Slemrod 2007, Fuest & Riedel 2009 and Slemrod & Weber 2012 for surveys). However, with the exception of the rare occasions when randomised audits are available (Slemrod *et al.* 2001; Kleven *et al.* 2011), methodological limitations mean that the credibility and precision of these estimates are questionable. Our paper contributes a novel methodology for the estimation of evasion using quasi-experimental variation created by tax policy. The approach generates robust estimates of evasion and can be easily replicated in other contexts as the tax variation needed is ubiquitous, especially in the developing world. Fourth, our paper is related to the literature on taxable income elasticities (Saez *et al.* 2012), and especially work that emphasizes the endogeneity of taxable income elasticities to the broadness of the tax base (Slemrod & Kopczuk 2002; Kopczuk 2005). Finally, we contribute to the large stream of literature studying responses by corporations to the tax code (see Auerbach 2002, Hassett & Hubbard 2002 and Auerbach *et al.* 2010 for surveys, and Gruber & Rauh 2007, Bach 2012, Dwenger & Steiner n.d., Kawano & Slemrod 2012 and Devereux *et al.* 2013 for recent estimates of the elasticity of corporate taxable income with respect to the effective marginal tax rate).

The paper is organized as follows. Section 2.2 presents our conceptual framework, which is used in section 2.3 to develop an empirical methodology based on minimum tax

[32] Both the empirical and the theoretical literatures on public finance and development are surveyed in Besley & Persson (2013).

schemes. Section 2.4 describes the context and data and section 2.5 presents our results. Section 2.6 numerically analyzes optimal policy, and section 2.7 concludes.

2.2 Conceptual Framework

This section introduces a stylized model of the firm to analyze the optimal design of a tax on the firm's activities. Our analysis focuses on the firm's responses to changes in the tax base and changes in the tax rate, both in an environment with and without tax evasion. 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 towards a distortionary tax on output if this discourages tax evasion by firms. The stylized model allows us to identify sufficient statistics that capture this trade-off between production efficiency and revenue efficiency (compliance) and guides our empirical strategy in the next section. Our analysis is only partial and ignores the general equilibrium impact of different tax policies. These effects are important for tax design, regardless of the presence of evasion, but will not by captured by our empirical strategy either.

2.2.1 Firm Behavior and Tax Policy Without Evasion

A firm chooses how much output y to produce at a convex cost $c(y)$. The firm pays taxes $T[y, c(y)] = \tau [y - \mu c(y)]$, which depends on the tax rate τ and a tax base parameter μ. The tax base parameter equals 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 to a pure profit tax base when increasing μ from 0 to 1. The firm's after-tax profits equal

$$\Pi(y) = (1-\tau) y - c(y) + \tau\mu c(y). \tag{2.2.1}$$

The profit-maximizing output level solves

$$c'(y) = 1 - \tau \frac{1-\mu}{1-\tau\mu} \equiv 1 - \omega, \tag{2.2.2}$$

where ω denotes the tax wedge between the social and private return to output. For a pure profit tax base ($\mu = 1$), the tax wedge disappears and the output choice is efficient, regardless of the tax rate. For an output tax base ($\mu = 0$), the tax wedge equals the tax rate. The impact of the tax rate τ and the base parameter μ on the firm's output choice depends on the implied change in the tax wedge ω, with $\frac{\partial\omega}{\partial\tau} \geq 0$ and $\frac{\partial\omega}{\partial\mu} \leq 0$. The change from a high tax rate on a profit tax base to a lower tax rate on a broader output tax base will only affect the firm's output choice if it affects the tax wedge ω.

The government sets 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. Hence, the welfare objective of the government can be written as

$$W = \Pi(y) + \lambda \{T[y, c(y)] - R\}, \tag{2.2.3}$$

where the firm's output choice satisfies (2.2.2) and $\lambda \geq 1$ denotes the (endogenous) marginal cost of public funds. The welfare effect of changing the tax parameters τ, μ can be decomposed into a *mechanical* welfare effect from transferring resources from the firm to the government for a given output level and a *behavioral* welfare effect due to the response in output. While the behavioral response in y affects welfare through government revenue, it has only a second-order welfare effect through firm profits (envelope result following from $\Pi'(y) = 0$). We may write the mechanical welfare effect (normalized by the marginal cost of funds λ) of the tax rate τ as $M_\tau \equiv [y - \mu c(y)] \times [\lambda - 1]/\lambda \geq 0$. We write the mechanical welfare effect (again normalized by λ) of the tax base parameter μ as $M_\mu \equiv -\tau c(y) \times [\lambda - 1]/\lambda \leq 0$. Both mechanical effects equal 0 if the marginal cost of public funds λ equals 1. The total welfare impact of τ and μ equal respectively

$$\begin{aligned} \frac{\partial W}{\partial \tau}/\lambda &= M_\tau + \omega \frac{\partial y}{\partial \tau}, && (2.2.4) \\ \frac{\partial W}{\partial \mu}/\lambda &= M_\mu + \omega \frac{\partial y}{\partial \mu}. && (2.2.5) \end{aligned}$$

This allows us to establish a natural implication of the production efficiency theorem in this stylized model. With a pure profit tax base, the government can raise taxes without distorting the firm's output. Hence, if possible, it is optimal to tax pure profits.

Proposition 1 (Production Efficiency). *With perfect tax enforcement, the optimal tax base is given by the firm's pure profit (i.e., $\mu = 1$).*

Proof. For $\mu = 1$, the government can increase tax revenues by increasing the tax rate without affecting the production choice. Hence, the marginal cost of public funds $\lambda = 1$. The government sets $\tau = R/[y - c(y)]$. For any $\mu < 1$, we can increase μ by $d\mu$ and increase τ by $d\tau = \frac{\tau c(y)}{y - \mu c(y)} d\mu$ so that the mechanical welfare effects cancel out. Hence, the impact on welfare equals

$$\begin{aligned} dW &= \frac{\partial W}{\partial \tau} d\tau + \frac{\partial W}{\partial \mu} d\mu \\ &= \lambda \omega \frac{\partial y}{\partial \omega} \frac{\partial \omega}{\partial \mu} \left[\frac{\partial \omega / \partial \tau}{\partial \omega / \partial \mu} \frac{\tau c(y)}{y - \mu c(y)} + 1 \right] d\mu. \end{aligned}$$

The impact on welfare is positive if the wedge ω decreases in response to the change. This is true if and only if

$$-\frac{\partial\omega/\partial\tau}{\partial\omega/\partial\mu} \leq \frac{y-\mu c(y)}{\tau c(y)}.$$

By implicit differentiation of (2.2.2), we find $-\frac{\partial\omega/\partial\tau}{\partial\omega/\partial\mu} = \frac{1-\mu}{\tau(1-\tau)}$. Hence, the condition simplifies to the after-tax profits being positive,

$$(1-\tau)(y-\mu c(y)) - (1-\mu)c(y) \geq 0,$$

which is always satisfied for the firm's output choice. □

2.2.2 Firm Behavior and Tax Policy With Evasion

The previous section assumed that the government can perfectly enforce taxation, independently of the tax base and tax rate. We now relax this assumption and analyze optimal tax policy when the government's tax capacity is limited. In particular, we capture in our model the notion that an output tax is harder to evade than a profit tax, the argument being that it is harder to evade a broader base and possibly also that the *level* of output is more visible than the *difference* between output and input. We capture the relative ease of evading profit taxes by allowing firms to declare costs $c \neq \hat{c}(y)$ at a convex cost of misreporting $g(\hat{c}-c(y))$ with $g(0)=0$. The key implications are similar if we also allow the output level y to be misreported, as long as it remains harder to evade an output tax than a profit tax.

The firm again maximizes its after-tax profit, but now chooses both real output y (at real costs $c(y)$) and reported costs $\hat{c}$ for tax purposes,

$$\Pi(y,\hat{c}) = (1-\tau)y - c(y) + \tau\mu\hat{c} - g(\hat{c}-c(y)). \tag{2.2.6}$$

At the firm's optimum,

$$\begin{aligned} c'(y) &= 1-\omega, && (2.2.7)\\ g'(\hat{c}-c(y)) &= \tau\mu. && (2.2.8) \end{aligned}$$

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 $g(.)$ depends on the difference between reported and true costs rather than on the difference between reported and true tax liability (Allingham & Sandmo 1972; Yitzhaki 1974), but this assumption is not key for the main analytical insights that we present below. The output level depends on the tax wedge ω in exactly the same way as before, and is therefore not affected by the presence of evasion.[33]

With evasion, the government's tax revenue can be decomposed into the revenue based on the true tax base and the foregone revenue due to misreporting the base,

$$T\left[y,\hat{c}\right] = \tau \times \underbrace{\left[y - \mu\hat{c}\right]}_{\text{reported base}} = \tau \times \{\underbrace{\left[y - \mu c\left(y\right)\right]}_{\text{true base}} - \underbrace{\mu\left[\hat{c} - c\left(y\right)\right]}_{\text{unreported base}}\}.$$

The government's welfare objective can now be written as $W = \Pi\left(y,\hat{c}\right) + \lambda\left\{T\left[y,\hat{c}\right] - R\right\}$, where the (normalized) mechanical welfare effects of τ and μ are given by $M_\tau \equiv \left[y - \mu\hat{c}\right] \times \left[\lambda - 1\right]/\lambda \geq 0$ and $M_\mu \equiv -\tau\hat{c} \times \left[\lambda - 1\right]/\lambda \leq 0$. Hence, the total welfare effects of τ, μ equal

$$\frac{\partial W}{\partial \tau}/\lambda = M_\tau + \omega\frac{\partial y}{\partial \tau} - \tau\mu\frac{\partial\left(\hat{c} - c\right)}{\partial \tau}, \tag{2.2.9}$$

$$\frac{\partial W}{\partial \mu}/\lambda = M_\mu + \omega\frac{\partial y}{\partial \mu} - \tau\mu\frac{\partial\left(\hat{c} - c\right)}{\partial \mu}. \tag{2.2.10}$$

Both an increase in the tax rate ($\tau \uparrow$) and an increase in the tax base ($\mu \downarrow$) entail a positive mechanical welfare effect, but a negative revenue effect through a decrease in the firm's real output. However, while an increase in the tax rate increases the level of misreporting, an increase in the tax base decreases the level of misreporting. We may state the following key proposition:

Proposition 2 (**Production Inefficiency**). *With imperfect tax enforcement, the optimal tax base is interior, i.e.,* $\mu \in (0,1)$*. The optimal tax system satisfies*

$$\frac{\tau}{1-\tau} \cdot \frac{\partial \omega\left(\mu\right)}{\partial \tau} = G(\mu) \cdot \frac{\varepsilon_{\hat{c}-c}}{\varepsilon_y}, \tag{2.2.11}$$

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* $\tau\mu$ *and* $\varepsilon_y \equiv \frac{\partial y}{\partial(1-\omega)}\frac{1-\omega}{y} \geq 0$ *is the elasticity of real output with respect to* $1-\omega$*. We have* $\frac{\partial\omega(\mu)}{\partial\tau} = \frac{1-\mu}{(1-\tau\mu)^2} \geq 0$ *which satisfies* $\frac{\partial\omega(0)}{\partial\tau} = 1$, $\frac{\partial\omega(1)}{\partial\tau} = 0$ *and is monotonically decreasing in* μ *whenever* $\tau \in \left[0, \frac{1}{2-\mu}\right]$*. We have* $G(\mu) \equiv \left[\hat{c} - c\left(y\right)\right]/\hat{\Pi}(y) \geq 0$ *where* $\hat{\Pi}\left(y\right) \equiv \left(1-\tau\right)y - \left(1 - \tau\mu\right)\hat{c}$ *are reported after-tax profits. This satisfies* $G(0) = 0$ *and is monotonically increasing in* μ*.*

Proof. 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, i.e., $\frac{\partial W}{\partial\mu}/\lambda = M_\mu - \mu\frac{\partial(\hat{c}-c)}{\partial\mu} < 0$. Notice that this result holds, even for $\lambda = 1$ and thus $M_\mu = 0$.

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 $\frac{\partial W}{\partial\mu}/\lambda =$

[33] The independence of real production and evasion relies on the assumption of additively separable evasion costs $g\left(.\right)$ that depend only on the evasion level $\hat{c} - c\left(y\right)$, independently of the real output level y. This independence simplifies the analysis without changing the main substance of our results.

$M_\mu + \tau \frac{\partial y}{\partial \mu} > 0$ 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 again an increase $d\mu$ in μ and $d\tau = \frac{\tau \hat{c}}{y - \mu \hat{c}} d\mu$ such that the mechanical welfare effects cancel out. The welfare effect through the change in y is like in the proof of Proposition 1. Hence,

$$\begin{aligned} dW/\lambda &= \omega \frac{\partial y}{\partial \omega}\left[\frac{\partial \omega}{\partial \tau}\frac{\tau \hat{c}}{y-\mu\hat{c}} + \frac{\partial \omega}{\partial \mu}\right] d\mu - \tau\mu \frac{\partial\left(\hat{c}-c\right)}{\partial \tau\mu}\left[\frac{\partial \tau\mu}{\partial \tau}\frac{\tau\hat{c}}{y-\mu\hat{c}} + \frac{\partial \tau\mu}{\partial \mu}\right] d\mu \\ &= \omega \frac{\partial y}{\partial \omega}\frac{\partial \omega}{\partial \tau}\left[\frac{\tau\hat{c}}{y-\mu\hat{c}} + \frac{\partial \omega/\partial \mu}{\partial \omega/\partial \tau}\right] d\mu - \tau\mu\frac{\partial\left(\hat{c}-c\right)}{\partial \tau\mu}\left[\mu\frac{\tau\hat{c}}{y-\mu\hat{c}} + \tau\right] d\mu \end{aligned}$$

Rewriting this in terms of elasticities, we find

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

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. □

Hence, in the presence of profit evasion, it is always optimal to introduce at least some production inefficiency by setting $\mu < 1$. To understand the optimal tax rule (2.2.11), note that the left-hand side $\frac{\tau}{1-\tau} \cdot \frac{\partial \omega(\mu)}{\partial \tau}$ reflects the effective marginal tax wedge on real production. This production wedge is equal to $\frac{\tau}{1-\tau}$ when $\mu = 0$, equal to zero when $\mu = 1$, and typically monotonically decreasing between those two extremes.[34] 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 $G\left(\mu\right)$, which is zero when $\mu = 0$ and monotonically 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 concern, and so it will be socially optimal to move close to a pure profit tax by setting $\mu \approx 1$ (such that $\frac{\tau}{1-\tau} \cdot \frac{\partial \omega(\mu)}{\partial \tau} \approx 0$). 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 towards the output tax by lowering μ, thereby simultaneously decreasing the evasion rate $G(\mu)$ and increasing the production wedge until formula (2.2.11) is satisfied.[35] The former case is arguably

[34] The cross-derivative of the production wedge with respect to the tax rate and base may switch signs such that the production wedge may be locally increasing in μ for $\mu < \frac{1}{2} - \tau$. Hence, this can only occur when the tax rate is at least 50 percent.

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

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 vs. weak tax capacity. Finally, note that the optimal tax formula (2.2.11) also identifies sufficient statistics for determining the optimal tax base and rate in our stylized framework, which we will study empirically.[36]

2.3 Empirical Methodology Using Minimum Tax Schemes

Using our conceptual 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 output/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 (non-standard) kink points in firms' choice sets, and that they produce differential quasi-experimental variation in the incentives for real production and compliance.

2.3.1 Minimum Tax Kink and Bunching (Without Evasion)

We first consider the baseline model without evasion. Firms report turnover y and costs $c\left(y\right)$ and pay the maximum of a profit tax ($\mu = 1, \tau_{\pi}$) and an output tax ($\mu = 0, \tau_y$) where $\tau_y < \tau_{\pi}$. That is,

$$T\left[y, c\left(y\right)\right] = \max\left\{\tau_{\pi}\left[y - c\left(y\right)\right], \tau_y y\right\}. \tag{2.3.1}$$

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

$$\tau_{\pi}\left[y - c\left(y\right)\right] = \tau_y y \quad \Leftrightarrow \quad p \equiv \frac{y - c\left(y\right)}{y} = \frac{\tau_y}{\tau_{\pi}}. \tag{2.3.2}$$

This implies a fixed cutoff τ_y/τ_{π} for the profit rate p (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 profit rate crosses the cutoff, the tax rate and tax base change discontinuously, but the tax liability (2.3.1) is continuous. Hence, this is a kink (a discontinuous change

[36] Our decomposition into real output and evasion elasticities is not in contradiction with the sufficiency of taxable income elasticities for welfare analysis (Feldstein 1995, 1999). It is possible to rewrite equation (2.2.11) 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.

in *marginal* tax incentives) as opposed to a notch (a discontinuous change in *total* tax liability), but a conceptually different type of kink to those explored in previous work (Saez *et al.* 2012; Chetty *et al.* 2011) due to the joint change in tax rate and tax base.[37] In the model without evasion, firms choose only real output based on the marginal return $1-\omega$, which changes from 1 (profit tax) to $1-\tau_y$ (output tax) at the kink.

Figure 2.1 illustrates how the minimum tax kink at τ_y/τ_π creates bunching in the distribution of profit rates. The dashed line represents the distribution of profit rates before the introduction of a minimum tax (i.e., under a profit tax). Assuming a smooth distribution of firm productivities (through heterogeneity in marginal cost functions $c'(.)$), this baseline distribution of profit rates is smooth and we denote it by $f_0(p)$. The introduction of a minimum tax (i.e., an output tax for $p \leq \tau_y/\tau_\pi$) reduces the marginal return to output from 1 to $1-\tau_y$ for firms initially below the cutoff. Those firms respond by reducing their output levels, which leads to an increase in their profit rates under decreasing returns to scale (i.e., marginal costs $c'(y)$ larger than average costs $c(y)/y$). This creates a right-shift in the profit rate distribution below the cutoff (with no change above the cutoff) and produces 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 2.1.[38]

Bunchers at the kink point τ_y/τ_π come from a continuous segment $[\tau_y/\tau_\pi - \Delta p, \tau_y/\tau_\pi]$ of the baseline distribution $f_0(p)$ absent the kink, where Δp denotes the profit rate response by the marginal bunching firm. Assuming that the kink is small, the total amount of bunching is given by $B = \Delta p \cdot f_0\left(\frac{\tau_y}{\tau_\pi}\right)$. Hence, based on estimates of excess bunching B and a counterfactual density at the kink $f_0\left(\frac{\tau_y}{\tau_\pi}\right)$, it is possible to infer the profit rate change Δp induced by the kink. In the model without evasion, this profit rate response is directly proportional to the real output elasticity. Assuming again that the minimum tax kink is small,[39] total differentiation yields

$$\Delta p = \left[\frac{c}{y} - c'(y)\right]\frac{dy}{y} \simeq \frac{\tau_y^2}{\tau_\pi}\varepsilon_y, \tag{2.3.3}$$

where we use that $c'(y) = 1$ and $\frac{c}{y} = 1-p \simeq 1-\frac{\tau_y}{\tau_\pi}$ in the vicinity of the cutoff. The output elasticity is defined as $\varepsilon_y \equiv \frac{dy/y}{d(1-\omega)/(1-\omega)}$ and we use that $\frac{d(1-\omega)}{1-\omega} = -\tau_y$ when crossing the kink.

[37] See Kleven & Waseem (2013) for further discussion of the conceptual distinction between kinks and notches.

[38] This analysis assumes decreasing returns to scale (which is a reasonable assumption for the short run, but probably not for the long run). Under constant returns to scale, the model without evasion predicts zero bunching at the minimum tax kink. This only strengthens our conclusion below that, once we allow for both real and evasion responses, bunching at minimum tax kinks tends to be driven mainly by evasion.

[39] The small-kink assumption is common in bunching studies and has the advantage of avoiding parametric specifications of the cost functions $c(.), g(.)$.

Based on equation (2.3.3), we note that large bunching (large Δp) will 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 small due to the fact that output/turnover is a very broad base (for example, τ_y is at most 1% 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 $1-\omega$, 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. The next section characterizes bunching responses in the model with evasion.

2.3.2 Minimum Tax Kink and Bunching (With Evasion)

We now turn to the model with evasion. Firms switch from profit to output taxation when the reported profit rate $\hat{p} = (y - \hat{c})/y$ falls below the cutoff τ_y/τ_π. This kink point is associated with a differential change in the incentives for real output and compliance. The marginal return to real output $1-\omega$ changes from 1 to $1-\tau_y$ when switching from profit to output taxation as in the previous model, 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 absent the minimum tax, the introduction of the minimum tax reduces real output (a loss of production efficiency) and increases compliance (gain in revenue efficiency). Assuming decreasing returns to scale, both effects increase the reported profit rate below the cutoff and produces bunching from below as shown in Figure 2.1.

Using the decomposition $d\hat{c} = d(\hat{c} - c) + dc$ and totally differentiating, we now obtain

$$\Delta\hat{p} = \left[\frac{\hat{c}}{y} - c'(y)\right]\frac{dy}{y} - \frac{d(\hat{c}-c)}{y} \simeq \frac{\tau_y^2}{\tau_\pi}\varepsilon_y - \frac{d(\hat{c}-c)}{y}, \tag{2.3.4}$$

where we again use $c'(y) = 1$ and $\frac{\hat{c}}{y} = 1 - \hat{p} \simeq 1 - \frac{\tau_y}{\tau_\pi}$. The bunching response $\Delta\hat{p}$ thus depends on both the real output response and the evasion response, but in very different ways. The real output response will be small under any potentially believable output elasticity (due to the scale factor τ_y^2/τ_π as described above), and so a large bunching response $\Delta\hat{p}$ must imply a large evasion response to the output tax. While we cannot separately estimate real output and evasion responses using only one minimum tax kink, equation (2.3.4) 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 *random* variation in the output tax rate τ_y applying to this scheme (giving us more than one observation of $\Delta\hat{p}$ for the same values of the output elasticity

ε_y and the evasion response $d(\hat{c} - c)$ under the randomness assumption), it would be possible to separately estimate the real and evasion responses.[40] Random variation in the profit tax rate τ_π is not as useful for separately estimating output and evasion responses, because the profit tax rate directly affects the evasion reponse $d(\hat{c} - c)$ to the minimum tax kink (and so does not give us additional observations of $\Delta\hat{p}$ for the same values of $d(\hat{c} - c)$).[41]

2.3.3 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 and a small turnover tax, the change in real incentives at the kink is minor, implying that substantial bunching provides evidence for evasion. However, the model above is extremely stylized, and so this section further analyzes the robustness of this argument.

Distortionary Profit Tax

The assumption that the full deductibility of costs causes the profit tax not to distort a firm's real decisions is stark and stands in sharp contrast to the large body of literature analyzing the effective marginal tax rate implied by corporate income taxation and its real impact on corporations (see Hassett & Hubbard 2002). However, relaxing this assumption only strengthens our conclusion that observed bunching must be driven overwhelmingly by evasion responses. Other things equal, the presence of real distortions due to the profit tax regime implies that the deterioration of real incentives when firms move from profit to turnover taxation will be smaller. 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 ($\Delta p \leq 0$) under non-increasing returns to scale and thus moves it *away* from the kink. Rewriting equation (2.3.4),

$$\Delta\hat{p} = \Delta p - d\left(\frac{\hat{c} - c}{y}\right), \tag{2.3.5}$$

we see clearly that also in the case of a distortionary profit tax (implying $\Delta p \leq 0$ other things equal) real responses cannot be responsible for bunching at the minimum tax kink (corresponding to $\Delta\hat{p} > 0$). Hence, our argument does not rely on the small change in the production incentives around the kink, but only on the production incentives not being much lower for the turnover tax (with a small τ_y) than for the profit tax such that Δp

[40] We do have time variation in τ_y in our empirical application, but such variation is not plausibly random and so we focus on the bounding approach to separately estimate evasion responses.

[41] We do have variation in τ_π in our empirical application that is arguably random. Even if this does not allow us to separately estimate output and evasion responses, it is still very useful for providing an additional identification check on our bunching strategy.

is small or negative. We conclude that if the effective marginal tax rate under the profit tax were positive (rather than zero), our estimate of the evasion response based on the decomposition in (2.3.4) would provide a lower bound.

Distortionary Turnover Tax: Cascading and Extensive Responses

As we argued above, a low turnover tax rate generates only small distortions to *firm-level* production incentives at the *intensive* margin. Importantly, this does not imply that the overall distortionary effect of turnover taxes is small. First, even a small turnover tax could cause significant production inefficiencies because of *cascading* through the production chain. Cascading effects of multiple-stage production can imply effective distortions far higher than statutory tax rates (Keen 2013). Crucially, though, such general equilibrium distortions affect the distribution of firms on both sides of the minimum tax kink and therefore cannot generate bunching at the kink. The absence of cascading effects in our bunching estimates is a great advantage for our ability to identify evasion responses to tax base switches, but cascading would still form a potentially important part of an overall welfare evaluation of minimum tax schemes.

Second, due to the breadth of the base, even a small turnover tax can create a large average tax rate on profits for firms with low (actual) profit rates, which could cause firms to respond along the *extensive margin*. Since the tax liability is continuous around the minimum tax kink, the switch between profit and turnover taxation does not create extensive responses in the vicinity of the kink, but there could be extensive responses further away from the kink. Conceptually, excess bunching is a measure of the intensive response and is therefore not affected by potential extensive responses, but such effects would be relevant for a broader policy evaluation.[42]

Output Evasion

The model can be extended to include output evasion whereby firms report output $\hat{y}$ which may differ from their true output y and face a convex cost of doing so. In this case, firms' profits are given by

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

[42] In particular, since profit rates vary across sectors due to differences in technology, a *uniform* turnover tax rate (as in our empirical application to Pakistan) creates differential average tax rates on profits across different sectors and therefore distorts the sectoral allocation of labour and capital. Such effects may call for sector-specific turnover tax rates depending on, for example, the average profit rate in each sector.

and the analog of equation (2.3.4) becomes

$$\begin{aligned}\Delta\hat{p} &= \left[\frac{\hat{c}}{\hat{y}} - c'\left(y\right)\right]\frac{dy}{\hat{y}} - \frac{d\left(\hat{c}-c\right)}{\hat{y}} - \frac{\hat{c}}{\hat{y}}\frac{d\left(y-\hat{y}\right)}{\hat{y}} \\ &\simeq \frac{\tau_y^2}{\tau_\pi}\varepsilon_y\frac{y}{\hat{y}} - \frac{d\left(\hat{c}-c\right)}{\hat{y}} - \left(1-\frac{\tau_y}{\tau_\pi}\right)\frac{d(y-\hat{y})}{\hat{y}} \qquad (2.3.6)\end{aligned}$$

decomposing the bunching response $\Delta\hat{p}$ into a real response in the first term and the two evasion responses. Given the lower tax rate ($\tau_y \ll \tau_\pi$), the incentives to underreport output are arguably smaller under the output tax than under the profit tax, which would further increase the bunching response. The key point to note is that this expression preserves the feature that the real output response in the first term will be small as it is scaled by τ_y^2/τ_π and so large bunching responses must reflect some combination of output and cost evasion responses. While we have done our welfare analysis in the presence of cost evasion only, the insights do not depend on the particular form evasion takes, as long as evasion is easier under a profit tax than under the output tax. Moreover, if it were true that the minimum tax makes it easier to misreport output, then firms would reduce their reported output more under the turnover tax, moving them *away* from the kink, so the presence of bunching also directly supports the notion that evasion is easier under a profit tax regime.[43]

Pricing Power

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

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

where $\rho\left(y\right)$ is the price the firm receives, which depends negatively on output y. In this model, the analog of equation (2.3.4) is

$$\begin{aligned}\Delta\hat{p} &= \left[\frac{\hat{c}}{\rho\left(y\right)y}\left(1-\sigma\right) - c'\left(y\right)\right]\frac{dy}{y} - \frac{d\left(\hat{c}-c\right)}{\rho\left(y\right)y} \\ &\simeq \left(1-\sigma\right)\frac{\tau_y^2}{\tau_\pi}\varepsilon_y - \frac{d\left(\hat{c}-c\right)}{\rho\left(y\right)y} \qquad (2.3.7)\end{aligned}$$

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 $\hat{c}/\rho\left(y\right)y = 1 - \tau_y/\tau_\pi$ and $c'\left(y\right) = \rho(y)\left(1-\sigma\right)$ at the kink. Firms now reduce their prices when increasing output. Hence, the more elastic the demand, the

[43] Here we have assumed that it is not possible for firms to misreport their output for the minimum tax without simultaneously misreporting their output for the profit tax. In the Pakistani context this is reasonable as firms report output once and this is used to calculate both the profit tax and the minimum tax liabilities.

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 (2.3.4) a lower bound.

2.4 Context and data

This section discusses the corporate tax regime in Pakistan and the administrative tax return data we use.

2.4.1 Corporate Taxation: Minimum Tax Regime

The corporate income tax is an important source of revenue in Pakistan and currently raises 2.5% of GDP, which comprises about 25% of all federal tax revenues. The tax is contributed by more than 20,000 corporations, which file their tax returns every year. However, there is evidence of large-scale non-compliance. A recent World Bank study reports that only about half of the corporations registered with the Securities and Exchange Commission are also registered with the Federal Board of Revenue (FBR), Pakistan's tax administration.[44] Among the latter group, only half of the firms file a tax return and only 37% of the filers report zero or negative taxable income.[45] The empirical evidence on the overall tax gap in Pakistan is limited: FBR reports an estimate of 45%, but does not provide information on the estimation.[46] The World Bank study suggests a tax gap that might be as high as 218% of actual corporate income tax payments. This estimation draws on an input-output model for a selected group of sectors. In light of the large-scale non-compliance, policy makers in Pakistan have devised a taxation scheme which ensures that every operational corporation pays some tax every year. The scheme, which has been in place since 1991, combines a conventional corporate income tax (profit tax) with a minimum tax payable on annual sales (turnover tax). Under this scheme, every firm calculates both tax liabilities and pays whichever liability is higher.[47]

The profit rate threshold which determines a firm's tax base is at the tax rate ratio $\frac{\tau_y}{\tau_\pi}$.

[44] See World Bank (2009).

[45] Authors' own calculation.

[46] See ?.

[47] The tax code also allows certain deductions from both tax liabilities, so that firms in effect compare net profit tax liability to the net output tax liablity. Firms are also allowed to carry forward the tax paid in excess of the profit tax liability and can adjust it against next year's liability to the extent that the net liability does not fall below the output tax liability for that year. Such adjustment, if not exhausted, can be carried forward for a further period of up to five years (three years in 2008 and 2009). In the data, we observe that only 1.3% of firms claim such carry forward, which indicates that firms are either unaware of this option or observe their profit tax liability net of carry forward drop below output tax liability, in which case carry forward cannot be claimed.

This ratio varies across different groups of firms and across time. First, Pakistan operates a reduced profit tax scheme for recently incorporated firms. All companies which register after June 2005, have no more than 250 employees, have annual sales not more than Rs. 250 million, and paid-up capital of not more than Rs. 25 million are eligible for a lower profit tax rate.[48] Second, both the high and low profit tax rate and the output tax rate undergo changes over time during the period considered in this study. Table 2.1 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 remained the same for the entire period under consideration. In the years we study, around 50 percent of the tax payers are on the minimum tax. The share of tax revenues from the minimum tax even increased above 50 percent in 2010 when the rate increased to 1%.

2.4.2 Data

Our study uses administrative data from FBR, covering the universe of corporate income tax returns for the years 2006-2010.[49] Since July 2007, electronic filing has been mandatory for all companies, and over 90% of the returns used in our study were filed electronically. Electronic filing ensures that the data has much less measurement error than what 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, which are sporadic and generally perceived to be ineffective.

Two aspects of the data are worth keeping in mind. First, our dataset 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 which do not file tax returns are non-operational. Second, as discussed before, the Pakistani economy is characterized by a large informal sector and our dataset includes only formal sector firms. However, even in a developed economy with high tax compliance, the largest firms generate the vast majority of corporate tax revenues. For example, in the UK the largest 1% of firms contribute 80% of corporate tax revenue.[50] The skewedness of the income distribution should be yet more striking in Pakistan, so that the tax compliance of small and informal firms has only a marginal effect on total tax revenue.

We limit our analysis to firms that report both profit and turnover, and either the

[48] Our empirical analysis takes into account that some of these requirements change during the period of study.

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

[50] HMRC 2007/08.

incorporation date or the profit tax liability, which are required for allocating firms to the high and low profit rate groups.[51] We also subject the data to a number of checks for internal consistency detailed in table A.4 of the appendix. Our final dataset contains 24,290 firm-year observations.

2.5 Empirical Results

This section presents the results of our empirical analysis, examing how firms respond to the minimum tax policy. We first present evidence that there is sharp bunching at the minimum tax kink of the form predicted by our analysis in sections 2.2 and 2.3, and that it is caused by the presence of the kink. We then analyze heterogeneity in bunching across different subsamples, showing that it is consistent with the predictions of the previous literature on tax evasion, and finally we use the observed bunching to estimate the magnitude of evasion responses.

2.5.1 Bunching at Minimum Tax Kinks

In section 2.3 we showed that in the presence of a minimum taxation system we should observe bunching by firms at a kink around a threshold profit rate (profits as a proportion of turnover) given by the ratio of the two tax rates, τ_y/τ_π. Figure 2.2 shows evidence that firms do indeed bunch around the 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 distribution of the reported profit rate (profits as percentage of turnover) in bins of approximately 0.2 percentage points.[52] Panel A shows the density for high-rate firms (facing a profit tax rate of 35%) 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 $\frac{\tau_y}{\tau_\pi} = \frac{0.5\%}{35\%} = 1.43\%$ (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% other than the presence of the minimum tax scheme, this represents compelling evidence of a behavioral response to the scheme. Notice also that there is a modest amount of "natural" excess mass around the zero-profit point (much smaller than at the kink point) as many firms generate very little income.[53]

[51] Table A.5 in the appendix compares the firms we lose to those we are able to use.

[52] The exact bin width is chosen to ensure that kink points are always located at bin centres. This requires us to slightly vary the bin width between panels A-C and panel D as the distance between the two kinks in panel D is different due to the higher turnover tax rate.

[53] A small number of firms in the data report *precisely* zero profits, which represents a form of "round-number bunching" as analyzed in detail by Kleven & Waseem (2013). To eliminate this effect driven by the salience of zero, the empirical distributions in Figure 2.2 excludes observations with $\pi = 0$, so the excess mass around zero is not driven by round-number bunching. Figure A.6 in the appendix reproduced panel A of figure 2.2, using the raw data before the consistency checks and without

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 exploiting 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% and therefore a minimum tax kink located at $\frac{\tau_y}{\tau_\pi} = \frac{0.5\%}{20\%} = 2.5\%$. Besides the different location of the kink, the size of the kink is smaller for low-rate firms in terms of evasion incentives (the variation in $\mu{\cdot}\tau$ at the kink point is smaller when τ_π is smaller) but not real incentives (the variation in $1 - \omega$ depends only on 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.[54] Overall, panel B strongly supports the interpretation that excess bunching around kink points is indeed driven by tax incentives, and not by spurious differences in the underlying profit rate distributions.

Panels C and D of figure 2.2 exploit time variation in the kink, focusing on the sample of high-rate firms. Panel C shows that excess bunching at 1.43% completely disappears in 2008 when the minimum tax regime (i.e., turnover tax below a profit rate of 1.43%) is removed. The 2008 density instead exhibits a larger mass of firms with profit rates between 0 and 1.43%. 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 bunching moves from 1.43% to 2.86% 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.

Taken together, the panels of figure 2.2 provide compelling evidence that firms respond to the incentives created by the minimum tax scheme, and 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.

2.5.2 Heterogeneity in Evasion Opportunities

Our theoretical framework suggests that bunching responses will be driven primarily by changes in evasion. We therefore expect that firms for which evasion is easier will respond more to the minimum tax kink and hence display greater bunching. To explore whether this is the case, we split the sample using indicators of evasion opportunities identified by

dropping firms with $\pi = 0$. Our results are robust to including firms reporting $\pi = 0$.

[54] The low-rate distribution is more noisy than the high-rate distribution because the former represents a much smaller fraction (about 16%) of the population of firms.

the previous tax evasion literature and compare the size of bunching responses in those subsamples. First, Kleven *et al.* (2009) develop an agency model in which firms with a large number of employees find it more difficult to sustain collusion on evasion, and Kumler *et al.* (2012) find evidence that supports this theory in Mexico, so we split the sample by firm size as proxied for by the wage bill and turnover.[55] Second, Gordon & Li (2009) provide a model where firms that rely on formal credit are more tax compliant (as this creates a paper trail that governments can observe), an argument that is consistent with the cross-country evidence in Bachas & Jensen (2013), so interest payments as a proportion of turnover is our second indicator of evasion opportunities. Third, firms selling to final consumers have more opportunity to evade than firms selling to other firms (due to the absence of a verifiable paper trail on the former), consistent with the experimental evidence for Chile by Pomeranz (2013), so we split the sample into retailers and non-retailers.

We focus on the group of high-rate firms, which represents most of the data and therefore gives us more power to detect heterogeneity. Panels A and B of figure 2.3 plot the density of the profit rate around the kink at 1.43%, splitting the sample equally by salary payments (scaled by turnover) and turnover, respectively. In accordance with the theory, we find that small firms respond more strongly to the kink. In both panels, the small-firm distribution exhibits a larger spike at the kink than the large-firm distribution. Panel C shows bunching for firms with more or less need for financial intermediation, as proxied for by reported interest payments as a fraction of turnover. As expected, the density for firms with below median interest payments exhibits larger excess bunching at the kink. Finally, panel D examines bunching by sector, dividing the firms into "retailers" (firms that sell at least partly to final consumers) and "non-retailers" (firms that sell exclusively to other firms). In line with the theory, there is larger bunching at the kink in the retailer subsample. We therefore conclude that the patterns of heterogeneity across firms in the bunching we observe at the kink are consistent with the predictions of the previous literature on the determinants of tax evasion and our argument that bunching is driven predominantly by evasion responses.

2.5.3 Estimating Evasion Responses Using Bunching

Having established that bunching is present as predicted by our theoretical analysis, and that it varies in accordance with the predictions of existing theories of tax evasion, this section presents estimates of excess bunching and translates them into estimates of the magnitude of evasion responses. Following Chetty *et al.* (2011), we estimate a counterfactual density—what the distribution would have looked like absent the kink—by fitting a flexible polynomial to the observed density, excluding observations in a range

[55] Unfortunately, the number of employees is not available in the tax returns.

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 p_j the (midpoint) profit rate in bin j, the counterfactual density is obtained from a regression of the following form

$$d_j = \sum_{i=0}^{q} \beta_i (p_j)^i + \sum_{i=p_L}^{p_U} \gamma_i \cdot \mathbf{1}[p_j = i] + \nu_j, \tag{2.5.1}$$

where q is the order of the polynomial and $[p_L, p_U]$ is the excluded range. The counterfactual density is estimated as the predicted values from (2.5.1) omitting the contribution of the dummies in the excluded range, i.e. $\hat{d}_j = \sum_{i=0}^{q} \hat{\beta}_i (p_j)^i$, and excess bunching is then estimated as the area between the observed and counterfactual densities in the excluded range, $\hat{B} = \sum_{j=p_L}^{p_U} \left(d_j - \hat{d}_j \right)$. Standard errors are bootstrapped by sampling from the estimated errors with replacement.

Figure 2.4 compares the empirical density distributions to estimated counterfactual distributions (smooth solid lines) for the four samples examined in figure 2.2: high-rate firms in 2006/07/09 in panel A, low-rate firms in 2006/07/09 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.[56] To better evaluate the estimated counterfactuals, each panel also shows the empirical distribution for a comparison sample in light grey (low-rate firms in panel A, high-rate firms in panel B, 2006/07/09 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, i.e. $b = \hat{B}/E(\hat{d}_j \mid j \in [p_L, p_U])$. 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 2006/07/09 ($b = 4.44\,(0.1)$) 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. Futhermore, excess bunching by high-rate firms is larger during the years 2006/07/09 than in year 2010 ($b = 2.05\,(0.2)$), possibly because optimization frictions prevent some firms from responding to the change in the location of the kink in the short run or because of improvements in enforcement.

Table 2.2 converts our bunching estimates into evasion responses using the methodology developed in section 2.3. As shown earlier, the amount of bunching translates to a profit

[56] The excluded range $[p_L, p_U]$ 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, two bins on either side in panels B and D. The order of the polynomial q is five (seven for 2008), chosen such as to optimize the fit. Table A.2 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.

rate response via the relationship $\Delta p = B/f_0 \left(\frac{\tau_y}{\tau_\pi}\right) \simeq b \times binwidth$,[57] and the profit rate response is in turn linked to the combination of real output and evasion responses via equation (2.3.4):

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

The table shows estimates of excess bunching b in column (1), the profit rate response Δp 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). 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 133 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.[58] 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, 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% to 66.7% of profits across the different populations, with high-rate firms in 2006/07/09 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.[59]

Finally, it should be noted that the evasion responses which we estimate using bunching at minimum tax kinks are responses to the *differential* evasion opportunities offered by

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

[58] For example, Gruber & 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, and transplanting it to the Pakistani context, with all the caveats that entails, would imply that $0.2 = \frac{dCTI}{d\omega}\frac{\omega}{dCTI} = \frac{\partial CTI/\partial y}{CTI/y}\varepsilon_y$ and so even a real output elasticity of 15 would require marginal taxable profits to be 1.33% of average taxable profits to be reconcilable with their estimate.

[59] We find that the estimates of the evasion rate response are very similar when assuming that only output can be evaded, using equation (2.3.6). The results are reported in table A.3.

profit and output taxation. The fact that we see such large bunching is *prima facie* evidence that it is much harder to evade output taxes than profit taxes, consistent with the motivation of the policy and our stylized conceptual framework. In the extreme case where output taxes offer zero evasion opportunity (as in our stylized model), our estimates would capture *total* tax evasion by firms in Pakistan. More realistically, if output taxes offer some scope for evasion as well, our estimates of evasion responses are lower bounds on the total evasion level by firms in Pakistan.[60]

2.6 Numerical Analysis of Optimal Tax Policy

This section links our empirical results to the stylized model introduced in section 2.2. The model characterizes the trade-off between production and revenue efficiency when setting the tax rate and tax base, while our empirical analysis identifies sufficient statistics allowing us to evaluate this trade-off in Pakistan. At the optimum, as shown in Proposition 2, the effective marginal tax wedge is equal to the inverse of the output elasticity scaled by the evasion rate and the evasion elasticity,

$$\frac{\tau}{1-\tau} \cdot \frac{\partial \omega \left(\mu\right)}{\partial \tau} = G(\mu) \cdot \frac{\varepsilon_{\hat{c}-c}}{\varepsilon_y}.$$

The right-hand side of this expression can be rewritten in terms of the evasion rate response we have estimated empirically, i.e.

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

using $d(\tau\mu)/(\tau\mu) = -1$ at the minimum tax kink. Evaluated for the high-rate firms who face a profit tax rate of $\tau_\pi = .35$, our estimate of the evasion rate response implies that the right-hand side equals 1.22 for $\varepsilon_y = .5$. This exceeds the left-hand side which is bounded from above by $\frac{\tau}{1-\tau} = \frac{.35}{.65} = .54$, implying that a pure profit tax is not optimal and that welfare could be unambiguously increased by broadening the base and decreasing the tax rate.[61]

To evaluate the optimal tax base and rate μ, τ, the optimal tax rule needs to be considered jointly with a revenue requirement. Alternatively, our estimates can shed light on

[60] This is consistent with the fact that our evasion estimates are lower than the existing tax gap estimates for Pakistan discussed in section 2.4. Those tax gaps were measured as a fraction of *true* tax liability, but can easily be converted into tax gaps as a fraction of *actual* taxes paid (corresponding to our estimates in Table 2.2), in which case they would be larger than 100%.

[61] Note that the real output elasticity would need to exceed 1 for the right-hand side to fall below the upper bound for the left hand side. Evaluated at the output tax base taxed at rate $\tau_y = .005$, the left-hand side equals $\frac{\tau}{1-\tau} = \frac{.005}{.95}$ which now exceeds the right-hand side having fallen to 0 for $\mu = 0$. While this suggests that welfare could be increased by the opposite changes, this conclusion would always hold regardless of the estimates of the evasion response rate, in line with Proposition 2.

the optimal base μ for a given tax rate τ, using only the optimality condition in Proposition 2 as re-written above. This crucially depends on how quickly the effective tax rate increases and the evasion rate response decreases as the tax system moves from profit taxation ($\mu = 1$) towards turnover taxation ($\mu = 0$). While our model determines exactly how the effective tax rate depends on the tax base, our bunching estimates for low-rate and high-rate firms may be combined to reveal how the evasion rate response changes with the evasion incentive $\tau\mu$.[62] Since the profits of the low-rate firms are taxed at $\tau_\pi = .20$, their evasion incentive $\tau\mu$ is equal to $\frac{.20}{.35} = 57\%$ of the evasion incentive of high-rate firms. The smaller reponse by the low-rate firms imply that the right-hand side of the above equation would fall from 1.22 to .34 when decreasing the tax base parameter from $\mu = 1$ to $\mu = .57$, using that the evasion incentive $\tau\mu$ is symmetric in the rate and the base. As our model implies that the right-hand side would decrease to 0 when moving all the way to turnover taxation, we have three points for the right-hand side as a function of μ. Extrapolating between these three points as shown in panel A of Figure 2.5, we find that the two sides of the optimal tax rule are equal for $\mu = .578$. This suggests that only about half of the costs should be deductible from the corporate tax base at a tax rate of 35 percent. While this exact number relies on the specific assumptions of our model and calibration, the result indicates that the full cost deductibility granted by profit taxation are far from optimal when accounting for evasion. This conclusion is robust to different tax rates and output elasticities. Panel B of Figure 2.5 shows all combinations of the tax rate and base for which the trade-off between production and revenue efficiency is optimized. The optimal tax base moves further towards the turnover tax base when decreasing the tax rate, with an optimal μ close to zero for a tax rate of .005, which is the minimum tax rate in Pakistan. The figure also illustrates that a higher output elasticity would move the optimal tax system closer to profit taxation for any tax rate, but still far from a pure profit tax base with full cost deductibility, while a lower output elasticity would move the system closer to turnover taxation.

As discussed earlier, the simple conceptual framework underlying this welfare analysis ignores potentially important effects for the economy as a whole (such as cascading effects). These should be incorporated before making firm policy recommendations on the use of production-inefficient turnover taxes in developing countries, and so our numerical analysis of optimal tax policy can only provide a basic illustration that such instruments could be potentially optimal in the presence of limited tax capacity.

[62] Here we make the assumption that the variation in the profit tax rate τ_π across firms can be viewed as exogenous.

2.7 Conclusion

In this paper we have studied the trade-off between preserving production efficiency and preventing the corrosion of revenues by evasion faced by governments with limited tax enforcement capacity. In contrast to models without evasion in which the optimal tax base is as close as possible to pure profits (preserving production efficiency), we have shown theoretically that in the presence of evasion, the optimal tax base sacrifices some production efficiency in order to curtail evasion levels. Our optimality conditions relate the optimal marginal tax wedge on real production to the elasticities of real and evasion responses to the tax wedge, sufficient statistics that we study empirically. We have also developed a novel empirical approach showing that minimum taxation schemes of a type that is common throughout the developing world can be used to estimate tight bounds on the evasion response to switching from profits to a broader turnover base. Exploiting the small change in real incentives and the large change in evasion incentives presented by the minimum tax scheme corporations face in Pakistan, we estimate that the switch from a profit tax to a turnover tax reduces evasion levels by 60-70% of corporate income. Linking these estimates back to our conceptual framework, our tax rule implies that the optimal tax system has a base that is far broader than profits—in Pakistan, as little as 50% of costs should be deductible when taxed at 35%.

It should be noted that our policy conclusions are based on a setting in which tax enforcement capacity is exogenously given, which is a reasonable approximation of the short-medium term environment of a developing country. In the longer run where tax capacity is endogenous to economic development (Kleven *et al.* 2009) and investments in state capacity (Besley & Persson 2011, 2013), the policy recommendations would obviously change. In fact, the large compliance gains to production-inefficient policies that we estimate shows the potentially large social returns to greater tax capacity.

Finally, since the minimum tax scheme generating the variation that is required for our methodology is ubiquitous throughout the world, this analysis could be replicated across the world, in particular for countries with different levels of tax capacity, shedding further light on the returns to investments in tax capacity and how evasion by firms responds to tax incentives.

Table 2.1: Tax Schedule

Year	**Tax Rates (%)**		
	Profit (high)	Profit (low)	Turnover
2006	35	20	0.5
2007	35	20	0.5
2008	35	20	0
2009	35	20	0.5
2010	35	25	1

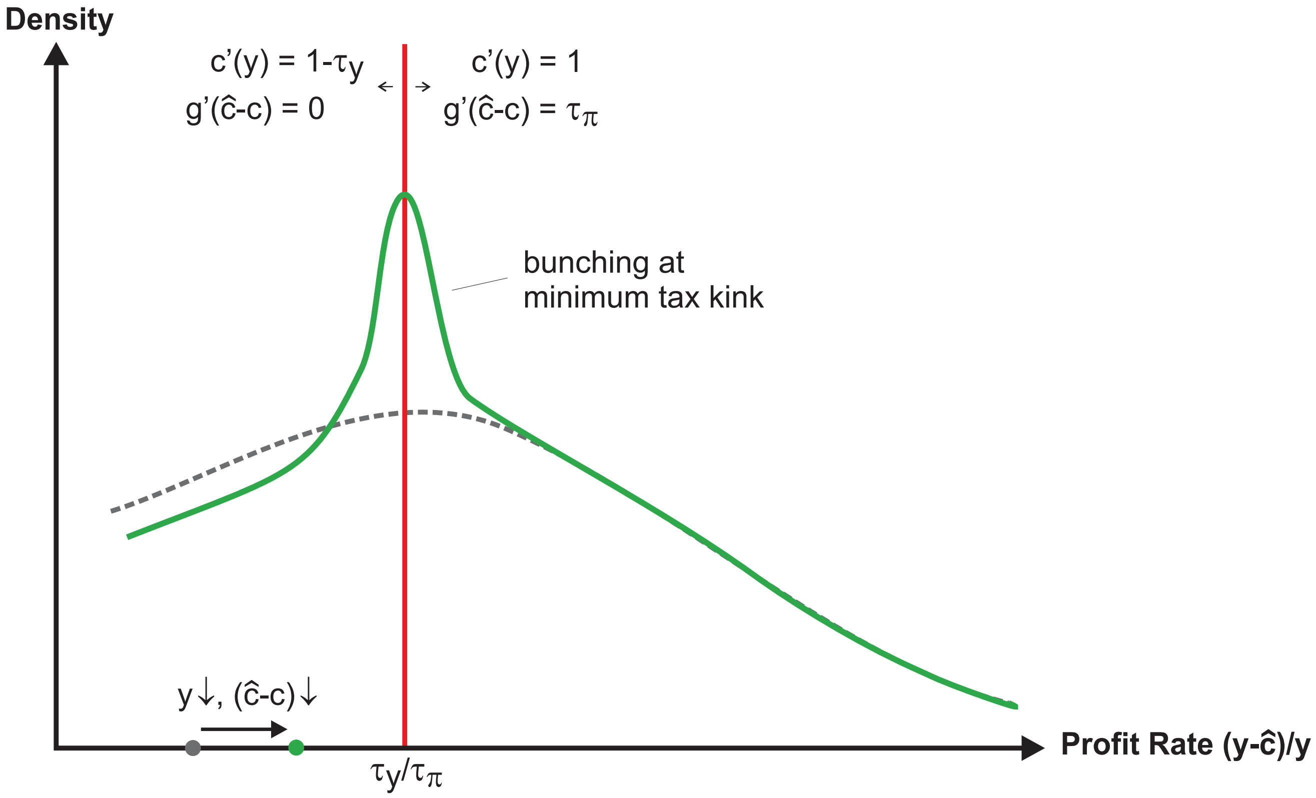
Notes: The table presents Pakistan's corporate income tax schedule for fiscal years 2006 to 2010. Fiscal year t runs from July 1 of year t to June 30 of year $t+1$. Profit rates are given in percentages. The low profit rate applies to firms which registered after June 2005, have no more than 250 employees, have annual sales of not more than Rs. 250 million, and paid-up capital of not more than Rs. 25 million. Our empirial analysis takes into account that the thresholds for some of these requirements change over time. All firms that do not meet these criteria are liable for the high profit rate. Firms calculate their net profit and turnover tax liablity (the tax code allows for specific deductions under each type of tax) and pay whichever liability is higher. Firms are allowed to carry forward the tax paid in excess of the profit tax liability and can adjust it against next year's liability to the extent that the net liability does not fall below the output tax liability for that year. Such adjustment, if not exhausted, can be carried forward for a further period of up to five years (three years in 2008 and 2009). In the data, we observe that only 1.3% of firms claim such carry forward, which indicates that firms are either unaware of this option or observe their profit tax liability net of carry forward drop below output tax liability, in which case carry forward cannot be claimed. We also exclude banks and financial firms, which face a standard tax rate of 38% in 2006 and 154 firms in sectors that were selectively given a lower turnover tax rate in 2010.

Table 2.2: Estimating Evasion Responses

	Observed Responses		**Without Evasion**	**With Evasion**			
	Bunching (b)	**Profit Rate (Δp)**	**Output Elasticity (ε_y)**	**Evasion Rate Response**			
				$\varepsilon_y = 0$	$\varepsilon_y = 0.5$	$\varepsilon_y = 1$	$\varepsilon_y = 5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
High-rate Firms, 2006/07/09	4.44	1.0	133.3	66.7	66.4	66.2	64.2
	(0.1)	(0.03)	(3.8)	(2.0)	(2.0)	(2.0)	(2.0)
Low-rate Firms, 2006/07/09	2.00	0.4	34.3	17.1	16.9	16.6	14.6
	(0.2)	(0.04)	(3.0)	(1.5)	(1.5)	(1.5)	(1.5)
High-rate Firms, 2010	2.05	0.4	14.7	14.7	14.2	13.7	9.7
	(0.2)	(0.03)	(1.2)	(1.2)	(1.2)	(1.2)	(1.2)

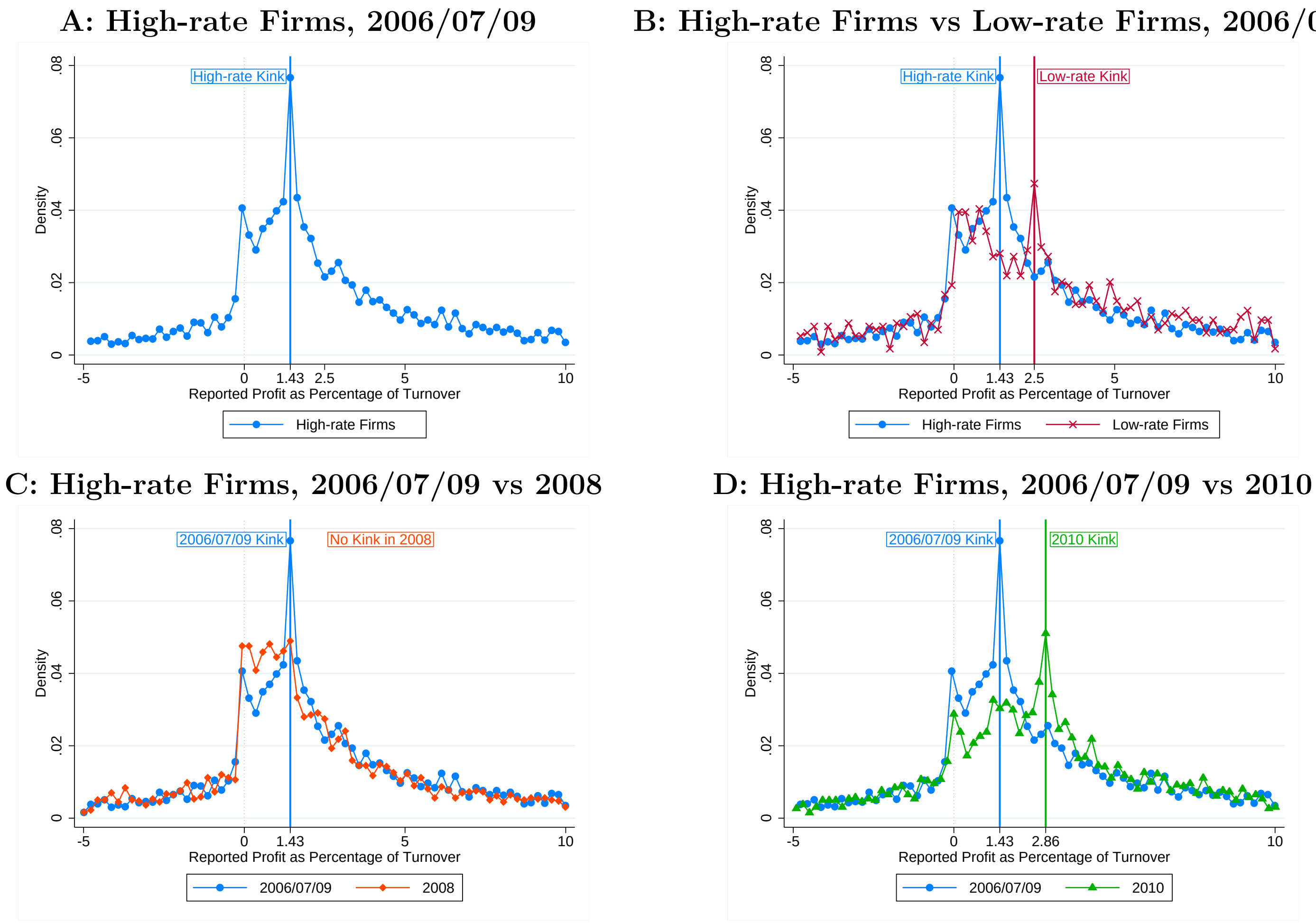
Notes: This table presents bunching and elasticity estimates for the subsamples considered in panels A, B and D of figure 2.4. Column (1) reproduces the bunching estimate b, based on estimating equation (2.5.1). Bunching b is the excess mass in the excluded range around the kink, in proportion to the average counterfactual density in the excluded range. Column (2) presents an estimate of the profit rate response associated with b, based on the relationship $\Delta p = B/f_0 \left(\tau_y/\tau_\pi\right) \simeq b \times binwidth$. 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. ε_y is estimated using the relationship $\Delta p = \left[c/y - c'\left(y\right)\right] dy/y \simeq \left(\tau_y^2/\tau_\pi\right) \varepsilon_y$. Columns (4)-(7) present estimates of the evasion response as 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 $\Delta \hat{p} = \left[\hat{c}/y - c'\left(y\right)\right] dy/y - \left[d\left(\hat{c} - c\right)/y\right] \simeq \left(\tau_y^2/\tau_\pi\right) \varepsilon_y - \left[d\left(\hat{c} - c\right)/y\right]$, assuming different real output elasticities $\varepsilon_y \in (0, 5)$. Bootstrapped standard errors are shown in parentheses.

Figure 2.1: Empirical Methodology



Notes: The figure illustrates the implications of the introduction of a minimum tax on the observed density distribution of reported profit rates $(y - \hat{c})/y$. The grey dashed line shows the smooth distribution of profit rates that would be observed in the absence of the minimum tax, while the green, solid line shows the distribution of profit rates that is observed in the presence of the minimum tax. As discussed in section 2.2.2, 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 towards the kink. Firms whose profit rate was close to the kink before the minimum tax was introduced pile up at the kink, which gives rise to an observed excess mass around the kink when accounting for optimization errors.

Figure 2.2: Bunching Evidence



A: High-rate Firms, 2006/07/09

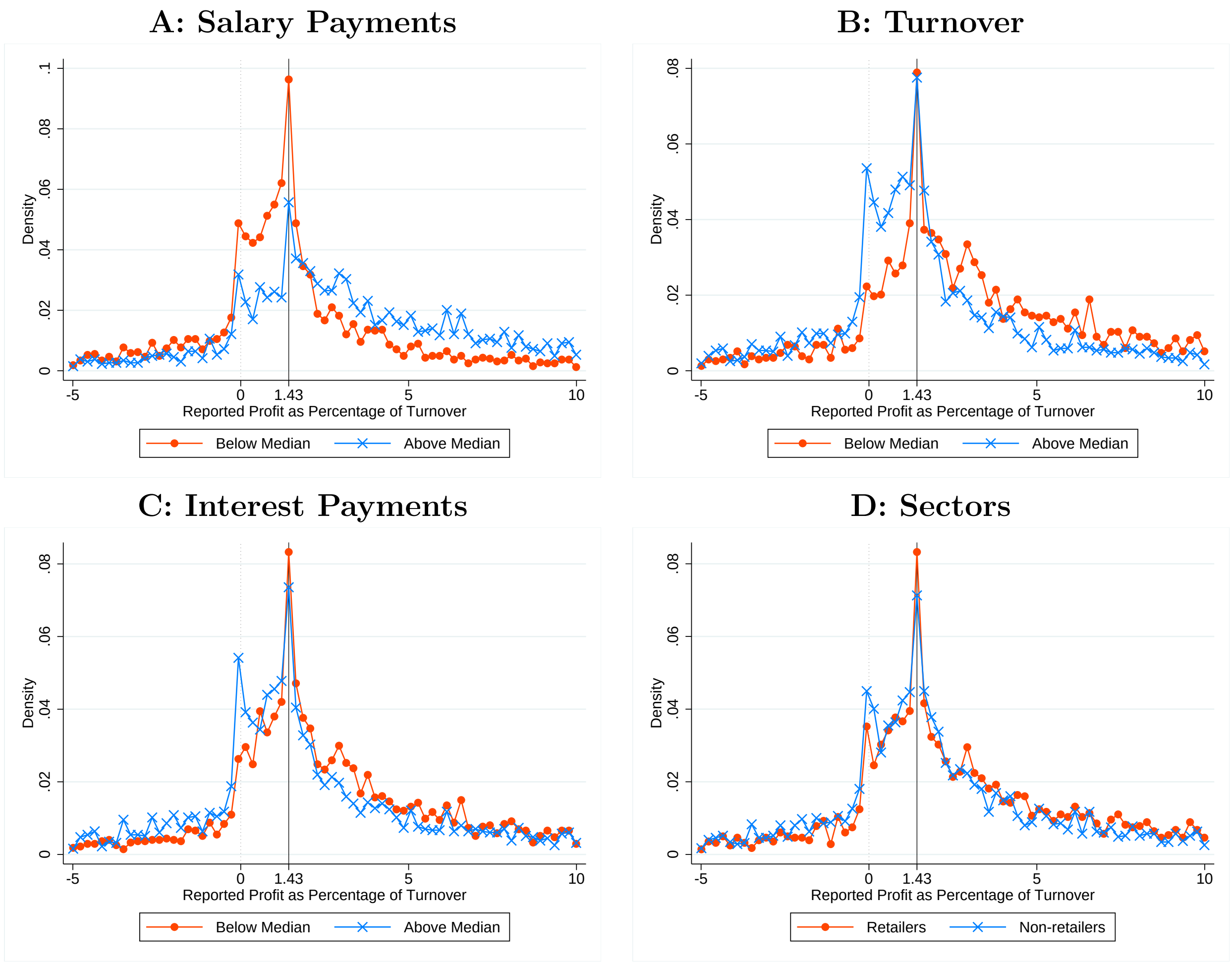
B: High-rate Firms vs Low-rate Firms, 2006/0

C: High-rate Firms, 2006/07/09 vs 2008

D: High-rate Firms, 2006/07/09 vs 2010

Notes: The figure shows the empirical density distribution of the profit rate (reported profit as percentage of turnover), for different groups of firms and time periods. The tax liablity for a firm with output y and cost $c(y)$ is $T\left[y,c\left(y\right)\right] = \max\left\{\tau_\pi\left[y-c\left(y\right)\right],\tau_y y\right\}$, where τ_π is the profit tax rate and τ_y is the turnover tax rate. The ratio of these two tax rates marks the kink at which firms move from the profit tax scheme (for profit rates above the kink) to the turnover tax scheme (for profit rates below the kink). For high-rate firms in 2006/07/09 (panel A), $\tau_\pi = 0.35$ and $\tau_y = 0.005$, placing the kink at a profit rate of 1.43%. For low-rate firms in 2006/07/09 (panel B), $\tau_\pi = 0.25$ and $\tau_y = 0.005$, placing the kink at a profit rate of 2.5%. For high rate firms in 2008 (panel C), the minimum tax scheme is abolished and the tax liablity is $T\left[y,c\left(y\right)\right] = \tau_\pi\left[y-c\left(y\right)\right]$ with $\tau_\pi = 0.35$, so that there is no kink. For high-rates firms in 2010 (panel D), $\tau_\pi = 0.35$ and $\tau_y = 0.01$, placing the kink at a profit rate of 2.86%. Kink points are marked by vertical solid lines, and the colour of the kink line matches the colour of the corresponding density. The zero profit point is marked by a vertical dotted line. The bin size is 0.214 (0.204 for 2010), chosen so that all kink points are bin centres.

Figure 2.3: Heterogeneity in Bunching



A: Salary Payments

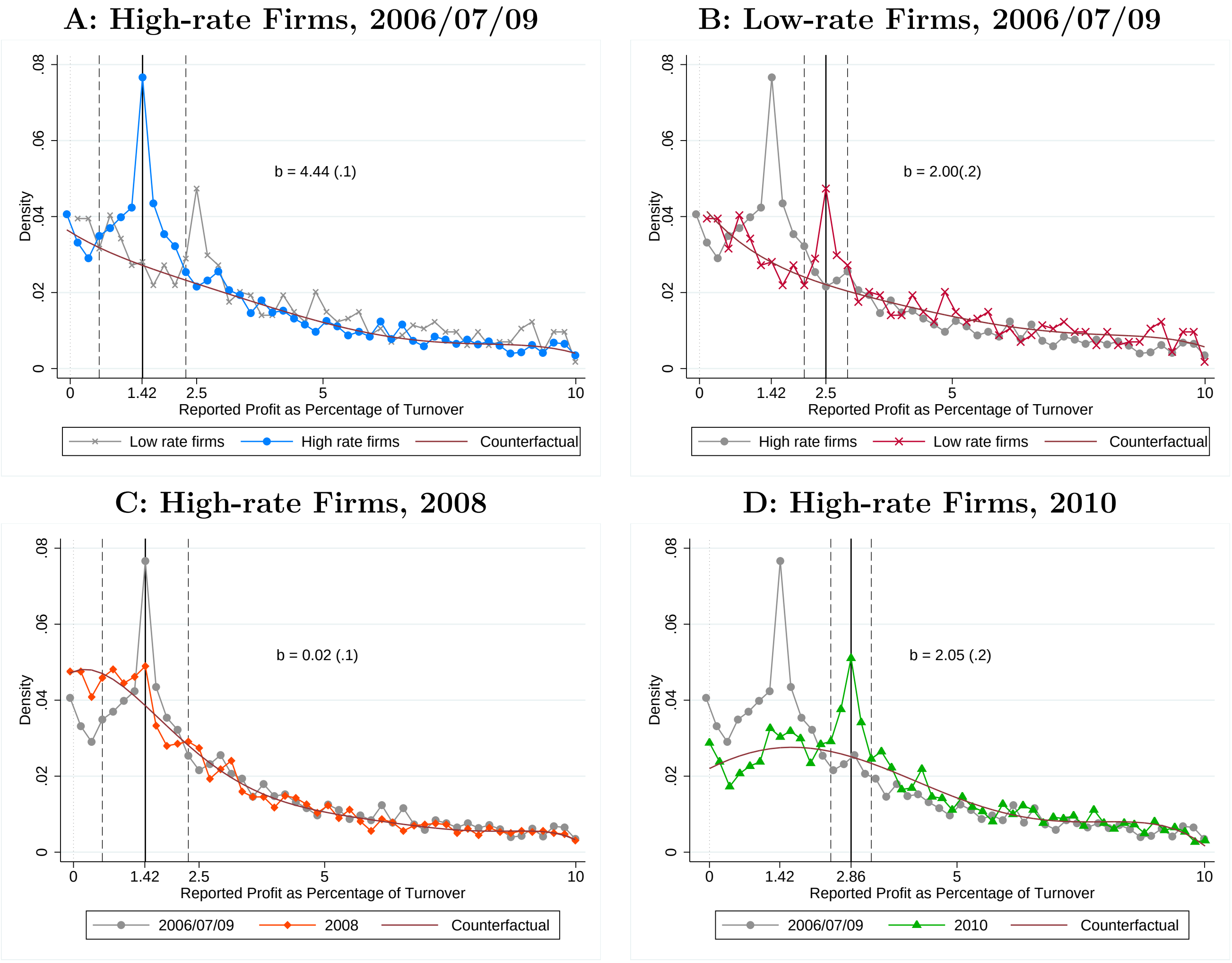
B: Turnover

C: Interest Payments

D: Sectors

Notes: The figure shows the empirical density distribution of the profit rate (reported profit as percentage of turnover), for different subsamples within the high-rate firms group in 2006/07/09. The tax liablity for a firm with output y and cost $c(y)$ is $T\left[y, c\left(y\right)\right] = \max\left\{\tau_{\pi}\left[y - c\left(y\right)\right], \tau_{y} y\right\}$, where τ_{π} is the profit tax rate and τ_{y} is the turnover tax rate. The ratio of these two tax rates marks the kink at which firms move from the profit tax scheme (for profit rates above the kink) to the turnover tax scheme (for profit rates below the kink). For high-rates firms in 2006/07/09, $\tau_{\pi} = 0.35$ and $\tau_{y} = 0.005$, placing the kink at a profit rate of 1.43%. In panels A-C, the high-rate firms sample is split by median salary payments, turnover and interest payements respectively. Salary payments and interest payments are scaled by turnover. The red (light) density is for firms below the median, and the blue (dark) is for firms above the median. Panel D splits the sample by sector, into retailers (red density) and non-retailers (blue density). The kink point is marked by a vertical solid line. The zero profit point is marked by a dotted line. The bin size in all panels is 0.214, chosen so that the kink point is a bin centre.

Figure 2.4: Estimating Evasion Responses



A: High-rate Firms, 2006/07/09

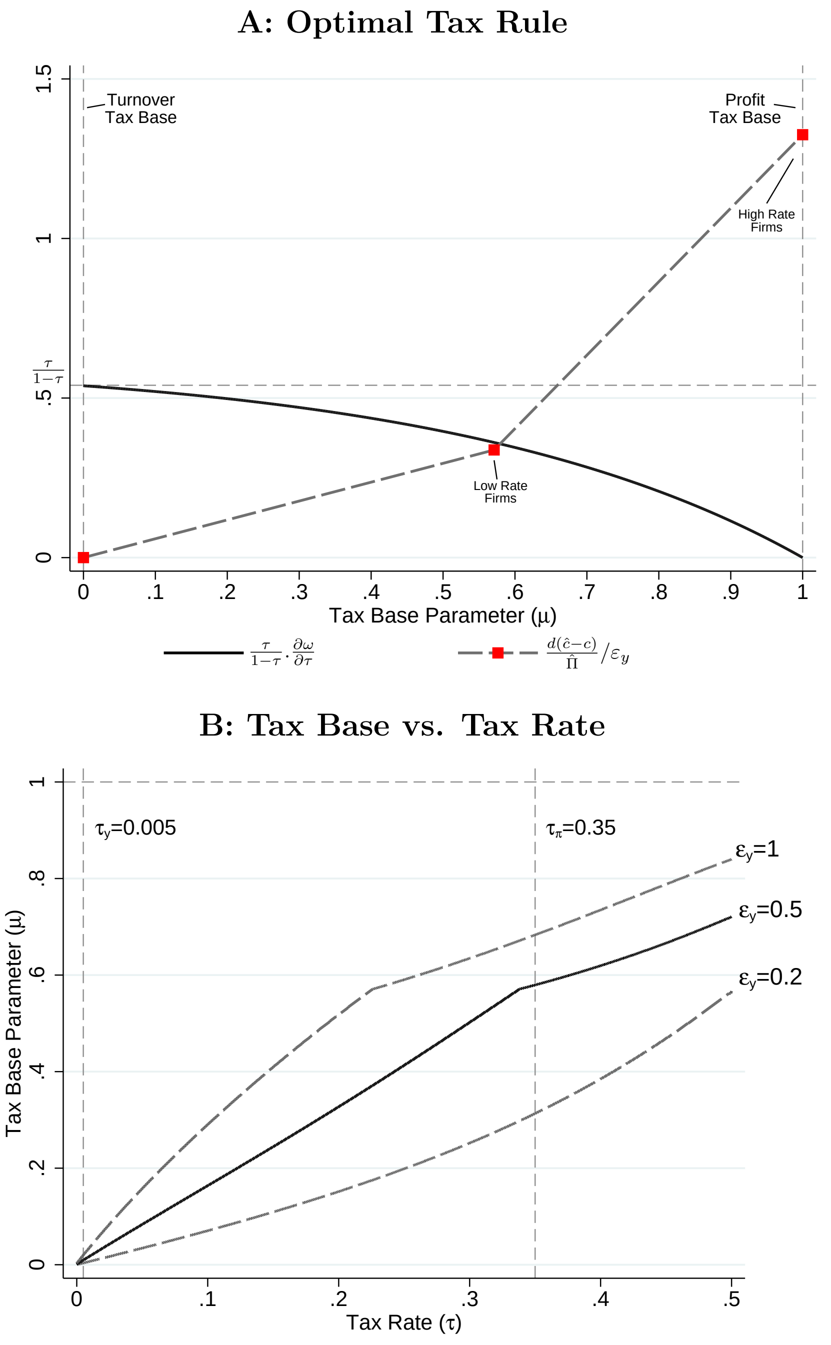
B: Low-rate Firms, 2006/07/09

C: High-rate Firms, 2008

D: High-rate Firms, 2010

Notes: The figure shows the empirical density distribution of the profit rate (reported profit as 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 2.2. The tax rate schedules and kink locations are explained in the footnotes to figure 2.2. The empirical counterfactual is the high-rate firms density for panels B, C and D, and the low-rate firms density 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 (2.5.1). 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 centres. 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.

Figure 2.5: Numerical Analysis of Optimal Tax Policy



A: Optimal Tax Rule

B: Tax Base vs. Tax Rate

Notes: The figure presents the numeric analysis of the optimal tax policy implied by the optimal tax rule (equation 2.2.11) and our empirical results. The solid black curve in panel A plots the left-hand side of the optimal tax rule equation as a function of μ for $\varepsilon_y = 0.5$ and $\tau_\pi = 0.35$. The three red markers on the dashed gray curve show respectively the right-hand side of the optimal tax rule at $\mu = 0$, at $\mu = {}^{0.2}/_{0.35}$ based on the evasion rate response estimated for the low-rate firms, and at $\mu = 1$ based on the evasion rate response estimated for the high-rate firms. By extrapolating between these three estimates, we find that the optimal tax base implied by the tax rule equals $\mu = 0.578$. In panel B we replicate the exercise to find the optimal tax base as a function of the tax rate for three different levels of the output elasticity.

Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan

3.1 Introduction

A central challenge in the literature on behavioral responses to taxes and transfers is how to estimate structural parameters when agents face optimization frictions such as switching costs, inattention, and inertia. Such frictions drive a wedge between the *structural* elasticity that matters for long-run welfare and the *observed* elasticity estimated from short-run variation in micro data (Chetty 2012). Most approaches in the literature ignore frictions, leading to downward-biased estimates of structural elasticities. Those that do account for frictions must do so either in a highly parametric setting or in a way that addresses the frictions only qualitatively. This chapter develops a framework for non-parametrically identifying optimization frictions and structural elasticities, and considers an application to income taxation in Pakistan.

Our framework exploits variation created by *notches* defined as discontinuities in the choice sets of individuals or firms. The specific focus is on notches that arise because incremental changes in earnings or labor supply cause discrete changes in the level of net tax liability, but the framework has a broader applicability than this. Notches are conceptually different from *kinks* defined as discontinuities in the *slope* of the choice set, as for example when the marginal tax rate jumps at bracket cutoffs in graduated income tax schedules. Although notches have received relatively little attention from economists, they are not uncommon in tax systems, welfare programs, social security, and regulation in many countries (Slemrod 2010).[63]

To understand the key idea of the chapter, consider a situation where income tax liability increases discretely at an earnings cutoff. Such a notch introduces an incentive for moving from a region above the cutoff to a point just below the cutoff, thereby creating a *hole* in the earnings distribution on the high-tax side and *excess bunching* in the

[63] Existing empirical studies have considered behavioral responses to notches in these various contexts, including the US Medicaid notch (Yelowitz 1995), social security notches (Gruber & Wise 2008; Manoli & Weber 2011), the US Saver's Credit notch (Ramnath 2013), the UK in-work benefit notch (Blundell & Hoynes 2004; Blundell & Shephard 2012), and car taxation notches (Sallee & Slemrod 2012).

earnings distribution on the low-tax side of the notch point.[64] What is particularly useful for empirical research is that the notch is associated with a region of strictly dominated choice above the cutoff where agents can increase both consumption and leisure by moving down below the cutoff. Intuitively, this occurs because the notch creates an implicit marginal tax rate of more than 100% over an interval. The dominated region should be completely empty in a frictionless world under any preferences, which implies that the observed density mass in this region can be used to measure attenuation bias from frictions. Therefore, by combining excess bunching below the notch (observed response attenuated by frictions) with the hole in the dominated region above the notch (frictions), it is possible to identify the structural elasticity that would govern behavior in the absence of frictions. Compared to recent bunching approaches using kinks (e.g. Saez 2010; Chetty *et al.* 2011), the conceptual advantage of notches relies on the possibility of using two moments of the density distribution to separately identify observed and structural elasticities. Compared to studies that address optimization frictions (e.g. Chetty *et al.* 2011; Chetty & Saez 2013), an additional advantage of notches is that they allow us to identify the sum total of all frictions while being agnostic about the specific sources of those frictions.

We apply our framework to the study of behavioral responses to income taxation in Pakistan. Despite the importance of understanding the link between tax policy and behavior in developing countries where fiscal capacity is limited, there is virtually no existing micro evidence from such settings.[65] Moreover, the issue of optimization frictions that is central to this chapter is likely to be at least as important in under-developed economies as in developed economies.

The Pakistani setting is chosen because it offers two important methodological advantages. First, the Pakistani income tax is designed as a piecewise linear schedule where each bracket is associated with a fixed *average* tax rate and therefore produces discontinuous jumps in tax liability at bracket cutoffs. These notches are substantial in size and therefore create very strong incentives for bunching below cutoffs and density holes above cutoffs. Second, we have gained access to administrative tax records covering the universe of personal income tax filers in Pakistan over the period 2006-2009. While the use of large administrative datasets is emerging as the norm for public finance research on developed countries, such data have so far been unavailable for research on developing countries. The combination of rich administrative data and sharp quasi-experimental variation from notches enables us to both demonstrate the potential of our method and to provide for the first time compelling evidence of behavioral responses to taxes for a developing economy.

Our main findings are the following. First, there is large and sharp excess bunching be-

[64] We use the intuitive term “hole” to describe the density distribution on the high-tax side of a notch point, but our framework shows that notches more generally create a triangular area of missing mass (that may not appear as a hole) between the observed and counterfactual (pre-notch) distributions.

[65] A recent survey of the literature on taxation and development is provided by **?**.

low every notch combined with missing mass ("holes") above every notch. Bunching and missing mass are much larger for self-employed individuals than for wage earners, consistent with the notion that self-employed individuals have more flexibility to adjust taxable income through tax evasion or real earnings. Second, even though observed bunching responses are large, those responses are strongly attenuated by optimization frictions as about 90 percent of wage earners and 50-80 percent of self-employed individuals located in strictly dominated regions are unresponsive to notches. This implies that, absent frictions, bunching would be 10 times larger than what we observe for wage earners and 2-5 times larger than what we observe for the self-employed. Third, while the combination of large observed bunching *and* large frictions implies that the taxable income response to notches would be extremely large absent frictions, the underlying structural elasticity driving this large response is relatively modest. The findings of large taxable income responses and small structural elasticities are not mutually inconsistent: notches create extremely strong distortions and therefore induce large behavioral responses even under small structural elasticities. Fourth, we present evidence on the dynamics and determinants of optimization frictions. Over time, the amount of dominated behavior (slowly) declines, so that the observed elasticity gets closer to the frictionless structural elasticity. This suggests that the estimated structural elasticities potentially represent long-run parameters.

3.2 Theory and Empirical Methodology

3.2.1 A Model of Behavioral Responses to Notches

We first analyze earnings responses to notches at the intensive margin, assuming a homogeneous structural earnings elasticity in the population, no optimization frictions, and a static setting. We subsequently consider generalizations that allow for heterogeneous elasticities, optimization frictions, dynamic aspects, and extensive responses.

Individual preferences are described by a quasi-linear and iso-elastic utility function

$$u = z - T\left(z\right) - \frac{n}{1+1/e} \cdot \left(\frac{z}{n}\right)^{1+1/e}, \tag{3.2.1}$$

where z is before-tax earnings, $T\left(z\right)$ is tax liability, and n is an ability parameter. This specification rules out income effects, but we discuss such effects below. As a baseline, we start by considering a linear tax system, $T\left(z\right) = t \cdot z$, where t is a proportional (average and marginal) tax rate. In this case, the maximization of utility with respect to earnings yields

$$z = n\left(1-t\right)^{e}, \tag{3.2.2}$$

where e is the elasticity of earnings with respect to the marginal net-of-tax rate $1-t$.

This is the structural parameter of interest as it serves as a sufficient statistic for tax revenue, welfare, and optimal taxation. At a zero tax rate, equation (3.2.2) implies $z = n$ and therefore the ability parameter can be interpreted as potential earnings. A positive tax rate depresses actual earnings below potential earnings, with the strength of the effect determined by the elasticity e.

There is a smooth distribution of ability in the population captured by a distribution function $F(n)$ and a density function $f(n)$. The combination of the ability distribution and the earnings supply function (3.2.2) yields an earnings distribution associated with the baseline linear tax system. We denote by $H_0(z)$, $h_0(z)$ the distribution and density functions for earnings associated with this baseline. Using (3.2.2), we obtain $H_0(z) = F\left(\frac{z}{(1-t)^e}\right)$ and hence $h_0(z) = H_0'(z) = f\left(\frac{z}{(1-t)^e}\right) / (1-t)^e$. Therefore, given a smooth tax system (no notches and no kinks), the smooth ability distribution converts into a smooth earnings distribution.

Suppose that a notch is introduced at the earnings cutoff z^*. This may be implemented as a discrete change in tax liability at the cutoff with no change in the marginal tax rate on either side (a "pure notch") or as a discrete change in the proportional tax rate at the cutoff (a "proportional tax notch"). The latter form combines a pure notch with a discrete change in the marginal tax rate (a kink). The empirical application considered below is based on proportional tax notches, but in this conceptual analysis we allow for pure notches as well. The notched tax schedule can be written as $T(z) = t \cdot z + [\Delta T + \Delta t \cdot z] \cdot 1(z > z^*)$ where ΔT is a pure notch, Δt is a proportional tax notch, and $1(.)$ is an indicator for being above the cutoff.

Figure 3.1 illustrates the implications of a proportional tax notch ($\Delta t > 0$, $\Delta T = 0$) in a budget set diagram (Panel A) and a density distribution diagram (Panel B). The notch creates a region of strictly dominated choice $\left(z^*, z^* + \Delta z^D\right]$ in which it is possible to increase both consumption and leisure by moving to the notch point z^*. There will be bunching at the notch point by all individuals who had incomes in an interval $(z^*, z^* + \Delta z^*]$ before the introduction of the notch, where the bunching interval is larger than the region of strictly dominated choice ($\Delta z^* > \Delta z^D$). Individual L has the lowest pre-notch income (lowest ability) among those who locate at the notch point; this individual chooses earnings z^* both before and after the tax change. Individual H has the highest pre-notch income (highest ability) among those who locate at the notch point; this individual chooses earnings $z^* + \Delta z^*$ before the tax change and is exactly indifferent between the notch point z^* and the interior point z^I after the tax change. Every individual between L and H locates at the notch point. There is a hole in the post-notch density distribution as no individual is willing to locate between z^* and z^I.[66]

[66] While the utility specification (3.2.1) eliminates income effects, the implication of such effects can be seen from Figure 3.1. The total response to the notch Δz^* can be divided into an uncompensated response $z^* + \Delta z^* - z^I$ (substitution + income effect) and a movement along the indifference curve $z^I - z^*$ (substitution effect). Earnings elasticities estimated from notches will in general be a mix

The basic idea in the empirical approach is that the width of the bunching segment Δz^* (corresponding to the earnings response of the marginal bunching individual) is determined by parameters of the tax notch and the elasticity e. Conversely, given knowledge of notch parameters and an estimate of the earnings response Δz^*, it is possible to uncover the elasticity e. To see this, consider the marginal bunching individual who is initially located at $z^* + \Delta z^*$ and whose ability level we denote by $n^* + \Delta n^*$. We exploit that this ability type is indifferent between the notch point z^* and the best interior point z^I. At the notch point z^*, the utility level is given by

$$u^N = (1-t)\, z^* - \frac{n^* + \Delta n^*}{1+1/e} \left(\frac{z^*}{n^* + \Delta n^*} \right)^{1+1/e}. \tag{3.2.3}$$

Using the first-order condition $z^I = (n^* + \Delta n^*)\,(1 - t - \Delta t)^e$, the utility level obtained at the best interior location can be written as

$$u^I = \left(\frac{1}{1+e} \right) (n^* + \Delta n^*)\,(1 - t - \Delta t)^{1+e} - \Delta T. \tag{3.2.4}$$

From the condition $u^N = u^I$ and using the the relationship $n^* + \Delta n^* = \frac{z^* + \Delta z^*}{(1-t)^e}$, we can rearrange terms so as to obtain

$$\frac{1}{1+\Delta z^*/z^*} \left[1 + \frac{\Delta T/z^*}{1-t} \right] - \frac{1}{1+1/e} \left[\frac{1}{1+\Delta z^*/z^*} \right]^{1+1/e} - \frac{1}{1+e} \left[1 - \frac{\Delta t}{1-t} \right]^{1+e} = 0. \tag{3.2.5}$$

This condition characterizes the relationship between the percentage earnings response $\frac{\Delta z^*}{z^*}$, the percentage change in the average net-of-tax rate created by each type of notch $\frac{\Delta T/z^*}{1-t}$, $\frac{\Delta t}{1-t}$, and the elasticity e. As we will directly estimate the earnings response Δz^* using bunching, it is useful to view the relationship (3.2.5) as defining the elasticity e as an implicit function of $\frac{\Delta z^*}{z^*}$, $\frac{\Delta T/z^*}{1-t}$, and $\frac{\Delta t}{1-t}$. It is not possible to obtain an explicit analytical solution for e, but it can be solved numerically given an estimate of Δz^* and observed values of the other arguments.

There are two important points to note about the elasticity formula (3.2.5). First, as the compensated elasticity e converges to zero (Leontief preferences), equation (3.2.5) implies

$$\lim_{e \to 0} \Delta z^* = \frac{\Delta T + \Delta t \cdot z^*}{1 - t - \Delta t} \equiv \Delta z^D. \tag{3.2.6}$$

Hence, under Leontief preferences, the bunching interval Δz^* converges to the strictly dominated range Δz^D in which taxpayers can increase both consumption and leisure by lowering earnings to the notch point.[67] The dominated range therefore represents a

of compensated and uncompensated elasticities, as is the case for elasticities estimated from *large* kinks (Saez 2010). The next section develops a reduced-form approach, which does not rely on the assumption of no income effects.

[67] The width of the dominated range Δz^D is defined such that the earnings level $z^* + \Delta z^D$ ensures the

lower bound on the earnings response to notches under any compensated elasticity in this frictionless model. The fact that notches create bunching even with a zero compensated elasticity represents a fundamental difference from kinks where a zero elasticity means zero bunching.

Second, although the preceding analysis considered a setting with only one notch, equation (3.2.5) encompasses settings with multiple notches. To see this, consider a situation with two cutoffs z_1^*, z_2^* associated with proportional tax notches $\Delta t_1, \Delta t_2$ and/or pure notches $\Delta T_1, \Delta T_2$. We may distinguish between two situations: (*i*) if the second notch is located outside the bunching segment of the first notch ($z_2^* \geq z_1^* + \Delta z_1^*$), then the two notches can be analyzed in isolation and the preceding analysis is unaffected. (*ii*) If the second notch is located inside the bunching segment of the first notch ($z_2^* < z_1^* + \Delta z_1^*$), then the marginal bunching individual at the first notch is coming from above the second notch. As above, an elasticity formula can be derived by exploiting that the marginal bunching individual must be indifferent between the notch point z_1^* and his best interior point z^I. It is necessary to distinguish between two different cases, which are illustrated in Figure A.8 of the appendix. If the best interior point is located in the top bracket ($z^I > z_2^*$), the elasticity formula is equivalent to equation (3.2.5) for $\Delta t \equiv \Delta t_1 + \Delta t_2$ and $\Delta T \equiv \Delta T_1 + \Delta T_2$. If the best interior point is instead located in the middle bracket ($z_1^* < z^I \leq z_2^*$), the elasticity formula is given by equation (3.2.5) for $\Delta t \equiv \Delta t_1$ and $\Delta T \equiv \Delta T_1$.[68] Section 3.2.3 describes how we deal empirically with the possibility of bunchers jumping multiple notches.

The determination of the elasticity e from equation (3.2.5) requires an estimate of the earnings response Δz^*. The model provides a relationship between the earnings response and estimable entities. Denoting excess bunching at the notch by B, we have

$$B = \int_{z^*}^{z^*+\Delta z^*} h_0(z)\, dz \approx h_0(z^*)\, \Delta z^*, \tag{3.2.7}$$

where the approximation assumes that the counterfactual density $h_0(z)$ is roughly constant on the bunching segment $(z^*, z^* + \Delta z^*)$. This approximation underlies existing bunching estimators, but we will account for potential curvature in the counterfactual density when estimating the earnings response from bunching.

We now consider the following extensions of the model: heterogeneity in elasticities, optimization frictions, dynamics, and extensive responses. Figure 3.2 illustrates the effect of a notch on the density distribution in the benchmark model (Panel A) and in various more general models (Panels B-D). To simplify the exposition, it is assumed that the notch is associated with a small change in the *marginal* tax rate above the cutoff, so that

same consumption as the notch point z^*, i.e. $(1 - t - \Delta t)\left(z^* + \Delta z^D\right) - \Delta T = (1 - t)\, z^*$.

[68] There is a third knife-edge case where the marginal buncher at the first notch is indifferent between the first and second notch points and where the latter is not a tangency point like z^I. In this case, the elasticity formula has to be modified.

intensive responses by those who stay above the notch can be ignored. In this case, the pre-notch and post-notch densities coincide above the bunching segment $(z^*, z^* + \Delta z^*)$.

Heterogeneity in Structural Elasticities

We allow for a joint distribution of abilities and elasticities represented by density $\tilde{f}(n, e)$ on the domain $(0, \infty) \times (0, \bar{e})$. At each elasticity level, behavioral responses can be characterized as in the benchmark model. The bunching segment at elasticity e is given by $(z^*, z^* + \Delta z_e^*)$, where Δz_e^* is increasing in e and takes the value Δz^D for $e = 0$. The post-notch earnings density in the full population will look like the solid blue curve in Panel B. The density is empty in the strictly dominated range and then increases gradually until it converges with the pre-notch density at $z^* + \Delta z_{\bar{e}}^*$. The grey shaded area in the post-notch density consists of those whose elasticity is too low for bunching given their location in the baseline earnings distribution.

With heterogeneity, bunching can be used to estimate the average earnings response $E[\Delta z_e^*]$. Denoting by $\tilde{h}_0(z, e)$ the joint earnings-elasticity distribution in the baseline without a notch and by $h_0(z) \equiv \int_e \tilde{h}_0(z, e)\, de$ the unconditional earnings distribution in the baseline, we have

$$B = \int_e \int_{z^*}^{z^* + \Delta z_e^*} \tilde{h}_0(z, e)\, dz de \approx h_0(z^*)\, E[\Delta z_e^*], \tag{3.2.8}$$

where the approximation again assumes that the counterfactual density is locally constant in earnings (but not elasticities). Using equation (3.2.8), estimates of excess bunching and the counterfactual earnings density reveal the average earnings response in the population.

Optimization Frictions

Optimization frictions such as adjustment costs and inattention have two potential implications. One is that individuals who would move to the notch point in the absence of frictions may stay above the notch. The other is that individuals who do respond may not be able to target the cutoff precisely, so that excess bunching manifests itself as diffuse excess mass rather than a point mass. In the empirical application, the first aspect turns out to be very important (there is significant density mass in strictly dominated ranges) while the second aspect is much less important (bunching is very sharp). This suggests a model where responding to the notch is associated with a fixed adjustment cost, but conditional on incurring the adjustment cost individuals are able to control income precisely. This is the situation depicted in Panel C where adjustment costs create additional mass on the bunching segment $(z^*, z^* + \Delta z_{\bar{e}}^*)$ compared to the frictionless model, but bunching still manifests itself as a sharp spike at the cutoff z^*. There is heterogeneity in adjustment costs, so that at each earnings-elasticity level some individuals respond and some do not. The light-grey area in the figure consists of those who do not respond because of low structural elasticities, while the dark-grey area consists of those who do not respond because of high adjustment costs.

A key distinction in this model is between the earnings response conditional on bunching Δz_e^* and the actual earnings response given frictions. We refer to the first one as the *structural* response (governed by the structural elasticity e) and the second one as the *observed* response (governed by the observed elasticity).[69] While existing micro studies generally capture observed elasticities attenuated by frictions, a central advantage of our notches framework is that it allows for a separate estimation of observed and structural elasticities. We describe two approaches which provide, respectively, lower and upper bounds on the structural elasticity.

For the first approach, we denote by $a(z,e)$ the share of individuals at earnings level z and elasticity e with sufficiently high adjustment costs that they are unresponsive to the notch. We then have

$$B = \int_e \int_{z^*}^{z^*+\Delta z_e^*} \left(1 - a(z,e)\right) \tilde{h}_0(z,e)\, dz de \approx h_0(z^*)\,(1-a^*)\, E\left[\Delta z_e^*\right], \tag{3.2.9}$$

where the approximation assumes a locally constant counterfactual density (as above) and a locally constant share of individuals with "large" adjustment costs, $a(z,e) = a^*$ for $z \in (z^*, z^* + \Delta z_e^*)$ and all e. In the above expression, $E[\Delta z_e^*]$ is the average structural response not affected by frictions while $(1-a^*)\, E[\Delta z_e^*]$ is the average observed response attenuated by frictions. Given estimates of $B, h_0(z^*)$, the two types of response can be separately identified using an estimate of the locally constant share a^* of individuals with large adjustment costs. This share can be estimated from the strictly dominated range where any remaining mass must be the result of frictions. Denoting by $h(z)$ the observed earnings density in the presence of the notch, we have $a^* \equiv \frac{\int_{z^*}^{z^*+\Delta z^D} h(z)dz}{\int_{z^*}^{z^*+\Delta z^D} h_0(z)dz}$.

Compared to existing bunching approaches, the innovation of our approach is to combine two moments of the distribution—bunching B and the hole in the dominated range $1-a^*$—to obtain a behavioral response not attenuated by frictions. From equation (3.2.9), the structural earnings response is proportional to $B/(1-a^*)$, which represents the amount of bunching that would materialize if individuals overcame adjustment costs. We use this inflated bunching measure to evaluate the structural elasticity e in equation (3.2.5). This implies that the larger is observed bunching *and* the smaller is the hole, the larger is the structural elasticity.

This approach arguably provides a *lower bound* on the structural elasticity. To see why, notice that $a(z,e)$ is an endogenous variable that depends on the utility gain of moving to the notch point and the distribution of adjustment costs. As the distance to the earnings cutoff increases, the utility gain of moving to the notch point falls and so the minimum adjustment cost preventing a response falls as well. If the distribution of adjustment costs is

[69] If optimization frictions disappear over long time horizons, the observed and structural elasticities reflect short-run and long-run elasticities, respectively.

smooth, this effect makes $a(z,e)$ increasing on the bunching segment $(z^*, z^*+\Delta z_e^*)$.[70] In this case, estimating $a(z,e)=a^*$ from the dominated range understates average frictions and therefore the structural elasticity. If the distribution of adjustment costs is discrete, this effect is weaker and the downward bias therefore smaller. In the extreme situation with dichotomous adjustment costs (zero or prohibitively high) such that fixed shares of the population either do or do not respond, the approach yields unbiased estimates of frictions and structural responses.

We consider a second approach that provides an *upper bound* on the structural elasticity. For this approach, note first that an exact measure of attenuation bias from frictions requires us to know how much of the observed mass on the bunching segment $(z^*, z^*+\Delta z_{\bar{e}}^*)$ can be explained by low elasticities in a frictionless world (light-grey area in Panel C of Figure 3.2). An extreme assumption is that none of it can be explained by low elasticities and that it is therefore all driven by frictions. This corresponds to an assumption of homogeneous structural elasticities at $e=\bar{e}$. In this case, the structural response can be determined as the point of convergence between the observed and counterfactual distributions. If there is heterogeneity in elasticities, this approach estimates the structural response by the highest-elasticity individuals and therefore represents an upper bound on the average structural response in the population. In the empirical application, we consider both the upper-bound approach ("convergence method") and the lower-bound approach ("bunching-hole method").

Finally, it will be useful for empirical applications to consider more carefully what the model implies about the shape of the post-notch distribution. In Panel C of Figure 3.2, the post-notch density is increasing on the bunching segment and features a real hole, but this is not a general prediction of the model. What is a more general prediction is that the area of missing mass above the notch point is *triangular*. This is because, for a given elasticity e, the utility gain of moving to the notch point is monotonically decreasing in earnings $z > z^*$ and converges to zero at $z = z^* + \Delta z_e^*$. Therefore, unless frictions are strongly negatively correlated with earnings and/or if elasticities are strongly positively correlated with earnings, the height of the missing mass area declines monotonically as we move to the right. Given a missing mass triangle, the shape of the post-notch (observed) distribution simply reflects the shape of the pre-notch (counterfactual) distribution. Figure A.9 in the online appendix shows some examples. If the counterfactual density is increasing or flat, the observed density will be increasing on the bunching segment and feature a hole. If the counterfactual density is weakly decreasing, the observed density will be flat or weakly increasing on the bunching segment and may not feature a hole. Finally, if the counterfactual density is strongly decreasing, the observed density will be

[70] If adjustment costs are negatively correlated with earnings, it is theoretically possible to overturn this effect. However, since the utility gain of moving to the notch point falls to zero over a relatively small earnings range, this would require an implausibly strong correlation between frictions and earnings.

decreasing above the notch and feature no hole.

Dynamics and Career Concerns

The preceding analysis extends to a dynamic setting with a few modifications. One modification is that bunching responses to a within-period (annual) tax schedule in a multi-period decision context may include intertemporal substitution. In that case, bunching relates to the Frisch elasticity instead of the static compensated elasticity (Saez 2010).

Another potential modification is in the characterization of the strictly dominated range. This modification is necessary only in dynamic frameworks where current earnings affect future wages through career concerns, learning by doing, etc. Assuming that the relationship between current earnings and future wages is continuous, the presence of career concerns reduces—but does not eliminate—the dominated range. This can be understood by considering the bounds of the static dominated range. Close to the lower bound z^*, current net-of-tax earnings are discretely lower than at the notch point while future net-of-tax earnings are only infinitesimally larger by continuity of the career effect. Given consumption smoothing behavior, this implies lower consumption in all periods along with lower leisure in the current period, so this is still strictly dominated. At the upper bound $z^*+\Delta z^D$, current net-of-tax earnings are the same as at the notch point while future net-of-tax earnings are discretely larger due to career effects. This allows a consumption smoothing individual to enjoy larger consumption in all periods (but less leisure in the current period), so this point is no longer strictly dominated. These arguments show that a strictly dominated range persists, but of a smaller width. The robustness of our method to dynamic career effects can therefore be checked by estimating a^* over smaller ranges $\left(z^*, z^* + \Delta z^D/K\right)$ where $K > 1$.[71]

Extensive Responses

A difference between notches and kinks is that the former, by introducing a discrete jump in tax liability, may create extensive responses. This includes real participation responses as well as movements between the formal and informal sectors. Our methodology is not designed to uncover extensive responses, but here we consider if such responses introduce bias in our estimates of intensive responses.

To see the implications of extensive responses, consider first a model with real participation responses and no adjustment costs. The analysis is extended to allow for informality and adjustment costs below. In the model, individuals choose earnings conditional on participation ($z > 0$), and then make a discrete choice between $z > 0$ and $z = 0$ facing a

[71] The preceding analysis potentially overstates the implications of career effects for the dominated range by implicitly assuming that the career effect is triggered by higher current *earnings* as opposed to just higher current *working hours* (corresponding to pure learning by doing). In the latter case, a worker with earnings at the cutoff z^* has the option of increasing hours worked (to reap the learning-by-doing benefit) without receiving any instantaneous compensation. In this case, the dominated earnings range would be completely unaffected by the presence of dynamic career effects.

fixed cost of participation q that is smoothly distributed in the population. Extending the formulation (3.2.1), utility from participation is given by $u(z - T(z), z) - q$ while utility from non-participation is denoted by u_0. This implies that an individual participates iff $q \leq u(z - T(z), z) - u_0 \equiv \bar{q}$.

If a notch is introduced at z^*, this creates both intensive and extensive responses by those with $z > z^*$. However, extensive responses will be negligible *just* above the cutoff based on a revealed preference argument. Consider individuals initially located at $z = z^* + \epsilon$ where $\epsilon > 0$ is sufficiently small that the cutoff z^* is preferred to the initial location. Such individuals respond either by moving to $z = z^*$ (intensive response) or by moving to $z = 0$ (extensive response), with the extensive response being preferred for those who were initially close to the indifference point between participation and non-participation. Denoting by $\bar{q}_0$ the threshold fixed cost under the baseline linear tax system, $T(z) = t \cdot z$, there will be extensive responses for those with $q \in (\bar{q}_0 - \Delta\bar{q}, \bar{q}_0)$ where

$$\Delta\bar{q} = u\left(\left(z^* + \epsilon\right)(1 - t), z^* + \epsilon\right) - u\left(z^*(1 - t), z^*\right), \tag{3.2.10}$$

in which we have used that the optimal point under the notched schedule (the cutoff z^*) avoids the notch. The above expression implies $\lim_{\epsilon \to 0} \Delta\bar{q} = 0$, so that there no extensive responses close to the cutoff. This is a very intuitive result: if in the absence of the notch an individual prefers earnings slightly above z^*, then in the presence of the notch he is better off moving to z^* (which is almost as good as the pre-notch situation) than moving to $z = 0$. It is straightforward to extend this result to a model with informality responses instead of real participation responses.[72] Moreover, the argument carries over to the case with adjustment costs as long as the (small) intensive response does not involve a *strictly larger* adjustment cost than the (large) extensive response, which is a mild assumption.[73] These results imply that extensive responses affect the density distribution as illustrated in Panel D of Figure 3.2.

These conceptual insights are very important for the empirical usefulness of notches. The fact that extensive responses do not occur locally around notches while intensive (bunching) responses occur only locally allows us to separate the two responses. In par-

[72] Consider a model in which individuals choose between earning z formally (paying taxes $T(z)$) or informally (paying zero taxes). There is a cost of informality q_I (capturing, for example, expected fines, moral costs, productivity losses of operating in cash, etc.) that is smoothly distributed in the population. The presence of informality costs ensures that informality is not always a strictly preferred choice (such that there is a formal sector in equilibrium). Utility under formality is given by $u(z - T(z), z)$ while utility under informality is given by $u(z, z) - q_I$, and hence an individual opts for formality iff $q_I \geq u(z, z) - u(z - T(z), z) \equiv \bar{q}_I$. From here, the argument that extensive (informality) responses do not occur in close proximity to the notch point z^* is analogous to the argument above.

[73] In a setting with informal production where the extensive response does not necessarily entail changing the level of real production, adjustment costs realistically arise because the informal worker has to adjust the production *process* to avoid getting detected with a very high probability. For example, a worker going informal must quit using banks and operate only in cash.

ticular, the bunching-hole method developed above exploits density mass in a narrow range below the cutoff relative to density mass in a narrow dominated range above the cutoff, and those local relative densities should not be substantially affected by extensive responses. On the other hand, the convergence method which relies on properties of the density distribution over a larger range *is* potentially sensitive to extensive responses. Section II.C describes how we deal with this issue.

3.2.2 A Reduced-Form Approximation of the Earnings Elasticity

The preceding analysis relies on a specific functional form for utility, and it would be useful to develop a reduced-form approach without such parametric reliance. A reduced-form method is less straightforward for notches than for kinks, because the behavioral response is driven by a jump in the average tax rate rather than a jump in the marginal tax rate of direct relevance to the structural parameter of interest. Here we set out a reduced-form approach for notches, which provides an approximation (upper bound) of the true structural elasticity.

The basic idea in the reduced-form approach is to relate the earnings response Δz^* to the change in the implicit marginal tax rate between z^* and $z^* + \Delta z^*$ created by the notch. Considering a proportional tax notch, the implicit marginal tax rate t^* is given by

$$t^* \equiv \frac{T\left(z^* + \Delta z^*\right) - T\left(z^*\right)}{\Delta z^*} = t + \frac{\Delta t \cdot \left(z^* + \Delta z^*\right)}{\Delta z^*} \approx t + \frac{\Delta t \cdot z^*}{\Delta z^*}, \qquad (3.2.11)$$

where the approximation requires that Δt is small (this approximation is not necessary, but simplifies slightly the elasticity formula below). The reduced-form elasticity of earnings with respect to the implicit net-of-tax rate is then defined as

$$e_R \equiv \frac{\Delta z^* / z^*}{\Delta t^* / \left(1 - t^*\right)} \approx \frac{\left(\Delta z^* / z^*\right)^2}{\Delta t / \left(1 - t\right)}. \qquad (3.2.12)$$

This simple quadratic formula provides an alternative to the parametric approach in the previous section. The formula essentially treats the notch as a hypothetical kink creating a jump in the marginal tax rate from t to t^*.

Figure 3.3 illustrates the relationship between the reduced-form and structural approaches using a budget set diagram. The reduced-form formula (3.2.12) treats the response to the notch Δz^* as if it were generated by the kink shown by the intersection of the lower budget segment with the solid green line. As shown in the figure, this kink schedule includes interior points that are strictly preferred to the cutoff by the individual initially located at $z^* + \Delta z^*$, who would therefore not become a buncher if faced with this kink. In this case, the bunching response to the notch Δz^* overstates the bunching response that would be created by the kink Δt^*, implying that the reduced-form elasticity e_R constitutes an upper bound. The key reason why this is true in the figure is that the best interior point z^I is located to the left—or at least not too far to the right—of $z^* + \Delta z^*$ in which case the marginal bunching individual under the notch would not be willing to bunch under the hypothetical kink. This corresponds to an assumption that the uncompensated earnings elasticity is not too strongly negative.[74]

3.2.3 Empirical Methodology and Identification

Our conceptual framework allows for the identification of structural parameters using excess bunching and missing mass in empirical density distributions around notches. Measures of bunching and missing mass will be based on a comparison between the empirical distribution and an estimated counterfactual distribution, using a procedure we now describe. We distinguish between a standard case with excess bunching only at notches and a case with excess bunching both at notches and round numbers (due to rounding in self-reported data).

Standard Case

Consider the (hypothetical) empirical density distribution in Panel A of Figure 3.4. The counterfactual density is estimated by fitting a flexible polynomial to the empirical density, excluding observations in a range $[z_L, z_U]$ around the notch point z^*. The excluded range should correspond to the area affected by bunching responses (area with excess bunching or missing mass), and we describe below how this is determined. Grouping individuals into small earnings bins indexed by j, the counterfactual distribution is obtained from a regression of the following form

$$c_j = \sum_{i=0}^{p} \beta_i \cdot (z_j)^i + \sum_{i=z_L}^{z_U} \gamma_i \cdot \mathbf{1}\left[z_j = i\right] + \nu_j, \tag{3.2.13}$$

where c_j is the number of individuals in bin j, z_j is the earnings level in bin j, and p is the order of the polynomial. The counterfactual distribution is estimated as the predicted values from (3.2.13) omitting the contribution of the dummies in the excluded range, i.e. $\hat{c}_j = \sum_{i=0}^{p} \hat{\beta}_i \cdot (z_j)^i$. Excess bunching and missing mass are estimated as the difference between the observed and counterfactual bin counts in the relevant earnings ranges, $\hat{B} = \sum_{j=z_L}^{z^*} (c_j - \hat{c}_j)$ and $\hat{M} = \sum_{j>z^*}^{z_U} (\hat{c}_j - c_j)$. The share of individuals in the dominated region D who are unresponsive is estimated as $\hat{a}^* = \sum_{j \in D} c_j / \sum_{j \in D} \hat{c}_j$. These estimates are illustrated in Panel B of Figure 3.4.

Standard errors are calculated using a bootstrap procedure in which we generate a large

[74] Given the size of the notch $\Delta t / (1-t)$ and a true functional form for utility, the bias of the reduced-form approach is determined by the percentage earnings response $\Delta z^* / z^*$. Figure A.10 in the appendix shows absolute and relative bias as a function of $\Delta z^* / z^*$, assuming that true preferences are quasi-linear as in (3.2.1). Absolute bias is increasing in $\Delta z^* / z^*$, but remains modest throughout a large range of responses. Relative bias is always largest at very small responses as $\Delta z^* = \Delta z^D$ implies $e = 0$ and $e_R > 0$.

number of earnings distributions (and associated estimates of each variable) by random resampling of residuals in (3.2.13). The standard error of each variable is defined as the standard deviation in the distribution of estimates of the given variable.

The approach relies on a credible determination of the excluded range $[z_L, z_U]$. Since excess bunching below a notch will typically be very sharp, the lower bound z_L can be determined visually without ambiguity. On the other hand, since missing mass above a notch is a more diffuse phenomenon occurring over a larger range, the upper bound z_U cannot be determined visually and a more disciplined approach is needed. We exploit that missing mass created by bunching responses must be equal to bunching mass, allowing us to pin down z_U by the condition $\hat{M} = \hat{B}$. To be precise, starting from a low initial value of the upper bound $z_U^0 \approx z^*$ and an initial estimate of the counterfactual $\hat{c}_j^0$ (with a flexible polynomial, we have $\hat{M}^0 \ll \hat{B}^0$), the upper bound is increased in small increments and the counterfactual re-estimated every time until we achieve $\hat{M}^k = \hat{B}^k$. The resulting estimate $\hat{z}_U = \hat{z}_U^k$ represents not just the upper bound of the excluded range and the area of missing mass, but is also the most natural definition of the "point of convergence" in the convergence method described earlier.[75]

We now address two potential concerns with our approach. First, if the notch creates extensive responses, this affects the observed distribution throughout the upper bracket in which case the estimated counterfactual (using observations above z_U) is not a "true" counterfactual stripped of *all* behavioral responses. This does not necessarily invalidate the estimation of the intensive elasticity, which requires us to estimate a "partial" counterfactual stripped of intensive responses only. Based on the theoretical model illustrated in Figure 3.2 (Panel D), intensive responses are concentrated in a triangular area close to the cutoff while extensive responses only become important further up. The idea of the estimation is to create a counterfactual by adding back the intensive-response triangle M, the total size of which must be equal to bunching mass B. Since we explicitly estimate z_U to ensure $\hat{M} = \hat{B}$, the only source of bias in z_U is functional form misspecification and we therefore carry out a sensitivity analysis with respect to the polynomial degree p. Moreover, as shown in the theory section, bias in z_U will have very little impact on the structural elasticity estimated from very *local* moments around the notch (B, a^*) as they should be roughly unaffected by extensive responses.

Second, the estimation procedure considers a single notch in isolation, but the empirical setting below consists of multiple notches. The presence of multiple notches is an issue only if bunchers are jumping more than one notch at a time. Although the conceptual

[75] By determining z_U such that $\hat{M} = \hat{B}$, we ignore a potential shift in the distribution within the interior of the upper bracket due to intensive responses by those who do not bunch. As can be seen in Figure 3.1 (Panel B), such a shift implies that bunching mass may not be fully matched by missing mass in a *small* region $(z^*, z_U]$, since some of the missing mass is spread over the entire distribution. This is a minor issue for notches associated with small changes in marginal incentives *within* the upper bracket $(z > z^*)$. This is satisfied for the empirical application below (and for many other notch settings as well).

framework allows us to deal with such scenarios, empirical implementation is difficult as bunching mass and missing mass are no longer matched at each notch separately. We therefore focus on notches sufficiently far apart that bunchers move only one notch. We make sure that this is satisfied by checking that $\hat{z}_U$ (estimated so that $\hat{M} = \hat{B}$) is significantly below the next notch point (for a large range of polynomial degrees p).

Identification in the Standard Case

It is useful to explicitly state the identifying assumptions necessary for notches to uncover structural elasticities. There are three key assumptions. (*i*) The counterfactual distribution is smooth such that excess bunching B identifies a behavioral response.[76] (*ii*) Bunchers come from a continuous set $M = B$ above the cutoff such that there exists a well-defined marginal buncher. (*iii*) The degree of friction a^* is locally constant and can therefore be inferred from the dominated region, allowing us pin down the frictionless behavioral response by the marginal buncher. While assumptions (*i*)-(*ii*) are quite weak, assumption (*iii*) is considerably stronger. Importantly, this set of assumptions is unambiguously weaker than the assumptions required for recent bunching approaches using kinks. Those approaches also require assumptions (*i*)-(*ii*) along with a much stronger third assumption (*iii'*) that the continuous set of movers M equals the *total* area under the counterfactual on a segment above the cutoff. This last assumption rules out any form of optimization friction, limiting the usefulness of kinks for the identification of structural parameters.

Round-Number Bunching

We find that taxpayers have a tendency to report taxable income in round numbers, which creates mass points at round numbers in the empirical distribution. We observe such rounding mainly for self-employed individuals (whose income is self-reported) and only to a very small extent for wage earners (whose income is mostly third-party reported), suggesting that this phenomenon is a side-effect of poor record keeping.

The anatomy of round-number bunching has a specific structure. First, some round numbers are rounder than others: for example, while there is excess mass at any income level that is a multiple of 1K, there is stronger excess mass at multiples of 5K, 10K, 25K and 50K. Second, there is rounding in both the annual and monthly dimension, the latter being a situation in which annual taxable income divided by 12 is a multiple of a round number. These two points together implies that round-number bunching is strongest at income levels that can be represented as multiples of many salient round numbers (1K, 5K, 10K, 25K, 50K,...) in both monthly and annual terms.

[76] This smoothness assumption also applies in the presence of extensive responses. In this case, the estimated counterfactual distribution is supposed to capture the distribution stripped of intensive responses, but not extensive responses. This "partial" counterfactual should also be smooth due to the fact that extensive responses to a notch do not affect the density *locally* around the cutoff (as shown in section II.A).

There are two conceptual points to note about round-number bunching. First, since notches are themselves located at salient round numbers, implementing the specification (3.2.13) without controlling for rounding would confound true notch bunching with round-number bunching and therefore overstate behavioral responses to the notch. Second, it is possible to control for round-number bunching at notches by using excess bunching at "similar round numbers" that are not notches as counterfactuals. In order to construct such round-number counterfactuals convincingly, we account for the underlying anatomy of rounding described above by estimating a rich set of round-number fixed effects that depend on the degree of roundness in both the annual and monthly dimension.

The regression specification we consider is the following

$$c_j = \sum_{i=0}^{q} \beta_i \cdot (z_j)^i + \sum_{r \in R, 12 \cdot R} \rho_r \cdot \mathbf{1}\left[\frac{z_j}{r} \in \mathbb{N}\right] + \sum_{i=z_L}^{z_U} \gamma_i \cdot \mathbf{1}\left[z_j = i\right] + \nu_j, \tag{3.2.14}$$

where $\mathbb{N}$ is the set of natural numbers, $R = \{1\text{K}, 5\text{K}, 10\text{K}, 25\text{K}, 50\text{K}\}$ is a vector of round-number multiples that capture annual rounding, and $12 \cdot R$ is a vector of round-number multiples that capture monthly rounding (as z_j is defined as annual income). The estimate of the counterfactual distribution is defined as the predicted values from the regression (3.2.14) omitting the contribution of the dummies around the notch, but not omitting the contribution of round-number dummies.

3.3 Application to Tax Notches in Pakistan

3.3.1 Income Tax and Enforcement System

The personal income tax in Pakistan currently raises revenue of 1.1 percent of GDP, or 11 percent of total tax revenue, and the share of registered taxpayers in the working-age population is less than 2 percent.[77] The low coverage of the income tax is consistent with the rest of the developing world. Individuals not registered for income tax fall in two categories: (*i*) those who are *legally* unregistered either because their income is below the exemption threshold or because of other types of exemptions (the most important of which is the exemption of agriculture income), (*ii*) those who are *illegally* unregistered and operate in the informal sector. Although informality is an important issue in Pakistan, the income exemption threshold (which is above the 80th percentile of the income distribution) and the exemption of agriculture (which represents about half of the workforce) can explain the bulk of non-registrations. Outside of the exemptions, the personal income tax applies to all wage earners, self-employed individuals and unincorporated firms. The tax schedule is fully individual-based and features a slightly higher exemption threshold for women than for men.

[77] See World Bank (2009).

What is crucial for our agenda is that the income tax is designed as a graduated schedule with a fixed *average* tax rate in each bracket and therefore a notch at each bracket cutoff. Figure 3.5 shows the average tax rate as a function of taxable income in Pakistani Rupees (PKR) for self-employed individuals (tax years 2006-09) and wage earners (tax years 2006-07).[78] We note the following about these schedules. First, the tax rate on self-employed individuals increases from 0 to 25 percent over thirteen notches, while the tax rate on wage earners increases from 0 to 20 percent over twenty notches (the first thirteen of which are included in the figure). Second, these notches create extremely strong incentives both because the average tax rate jumps are substantial *and* because they occur at high income levels. For example, at an income of PKR 500,000, one more rupee of income triggers tax liability of PKR 12,500 for the self-employed and PKR 5,000 for wage earners. Third, average tax rates are substantially higher for self-employed individuals than for wage earners, and the rule used to separate the two creates a different kind of notch. To be precise, each individual is classified as a self-employed individual (wage earner) if self-employment income as a share of total income is greater than or equal to (less than) 50%, and is then taxed according to the assigned schedule on the *entire* income. This creates a substantial income-composition notch at 50%, which we can use to estimate income shifting between wage income and self-employment income. Finally, tax schedules were fixed in nominal terms for self-employed individuals from 2006-2009 and for wage earners from 2006-2007 despite high inflation (8-20% annually). The wage earner schedule underwent a fundamental change in 2008, but we do not consider this reform here.[79]

Registered taxpayers are required to file income tax returns unless they meet certain filing exemption requirements.[80] The tax return is shown in Figure A.11 of the appendix.[81]

[78] Tax year t runs from July 1 of year t to June 30 of year $t+1$. During our data period (July 2006 to June 2010), the PKR-USD exchange rate was about 60 in the first half of the period and then increased to about 80 in the second half of the period.

[79] The 2008 reform for wage earners replaced the notch schedule by a complicated kink schedule. An earlier version of the paper analyzed this reform in detail, using it to confirm the identification strategy used here.

[80] In particular, wage earners are exempt from filing if (*i*) wage income is below 500K, (*ii*) the employer has filed a tax return (third-party report), and (*iii*) the taxpayer has no non-wage income. For such non-filers, taxable income is given by third-party reported wage income, which we observe in the data. Since filing is not costless, this exemption rule creates a *filing notch* for wage earners at 500K. Hence, behavioral responses to the 500K notch potentially conflate the effects of the tax rate and filing notches. However, a previous version of this paper exploits the 2008 reform for wage earners to separate the two effects and finds that the effect of the filing notch is small and statistically insignificant. We therefore ignore it in the empirical analysis below.

[81] The filed return is subject to a basic validation check by a computer software that uncovers any *internal* inconsistencies (e.g. between taxable income in cell 32 and tax liability in cell 33). Besides this validation check, the tax return is considered final unless selected for audit. Since our data represents pre-validation returns, inconsistencies between taxable income and tax liability may occur and provide a direct indicator of misperception/inattention from administrative data. This indicator captures misperception of either the tax rate schedule or the tax return itself (where the tax computation cells 33-41 create scope for confusion, especially for those subject to withholding). We exploit this unique measure of misperception in the empirical analysis.

The enforcement system involves some third-party reporting and withholding, the extent and form of which vary across taxpayer types. For most wage earners, there is third-party reporting and withholding by employers, a system known to deliver very strong enforcement in developed countries (Kleven *et al.* 2011). Self-employed individuals face no third-party reporting but are subject to certain withholding schemes. These schemes withhold taxes in connection with specific transactions (e.g. electricity bills, phone bills, and cash withdrawals), which are credited against income tax liability at the time of filing. This type of withholding comes with no third-party information on the tax base itself (taxable income), and is therefore not as powerful for enforcement as the system in place for wage earners. Tax evasion among self-employed individuals is therefore deterred primarily by the threat of audits and penalties, which tend to be infrequent and ineffective in Pakistan.

3.3.2 Data

Our study is based on administrative data from the Federal Board of Revenue (FBR) in Pakistan, including the universe of personal income tax returns filed for the tax years 2006-2009 (about 4 million observations in total). Returns were filed either electronically through the FBR website or by hard copy at designated bank branches and fed to computers using an IT firm distinct from FBR. This data collection process ensures that the data has much less measurement error than what is typically the case for developing countries. As far as we know, this is the first study to exploit such rich administrative tax data for a developing country.

The following aspects of the nature of the sample are worth keeping in mind. First, the universe of tax filers is not fully overlapping with the universe of registered taxpayers due to filing exemptions and potential non-compliance. Second, the population of tax filers is a high-income subsample of the general population due to the high income exemption threshold and the fact that larger incomes are more difficult to hide. Third, the population of tax filers is almost exclusively male (more than 99%), an implication of the individual tax system with a high exemption threshold combined with large gender inequality. Fourth, self-employment is much more prevalent among taxpayers in Pakistan (about half of the sample) than in developed countries. Finally, since our sample includes those who have selected into filing, they are likely to be a relatively tax-compliant subsample of the population.

3.3.3 Results for Self-Employed Individuals

This section presents empirical results for self-employed males.[82] As explained above, self-employed individuals have a tendency to report taxable income in round numbers,

[82] We drop the relatively small number of females as their tax schedule is slightly different at the bottom.

which creates round-number bunching in the empirical distribution and would lead to bias if ignored. We take a two-pronged approach to deal with rounding. First, we split the sample by those who report income in even thousands ("rounders") and those who do not ("non-rounders"). We separately analyze the continuous non-rounder sample (about 40% of filers), where we can implement the standard empirical specification (3.2.13).[83] Second, we consider the full sample of rounders and non-rounders, where we control for round-number bunching at notches using excess bunching at counterfactual round numbers that are not notches, using the empirical specification (3.2.14).

Figure 3.6 presents evidence from the first ten notches of the tax schedule for the non-rounder sample between 2006-2009. The top panels show the empirical distribution of taxable income around the six lower notches (Panel A) and the four upper notches (Panel B) as a histogram with dots at the upper bounds of each bin. Each notch point is demarcated by a vertical black line and is itself part of the tax-favored side of the notch. The following findings emerge from these panels. First, every notch is associated with large and sharp bunching just below the cutoff and missing mass above the cutoff, providing clear evidence of a response to the tax structure. Second, while the density falls discretely above notches and therefore features missing mass, there are no large holes in the distribution. This provides direct evidence of optimization frictions. Third, the shape of the distribution above notches is increasing at the bottom (where the surrounding distribution is increasing) and roughly flat at the middle and top (where the surrounding distribution is decreasing). This is consistent with the theory and suggests that the area of missing mass is triangular. Finally, the declining part of the empirical distribution features roughly a step-function pattern, a consequence of discrete drops at cutoffs and flatness in between notches.

The step-function pattern has two possible explanations. One possibility is that bunchers at a given notch are coming from the entire bracket above or even brackets higher up, so that the missing mass region (where the density is naturally flat) extends to the next notch or beyond. As explained in section 3.2.3, we investigate this possibility by estimating an upper bound of the missing mass region such that missing mass equals bunching mass given a smooth counterfactual distribution. We find that bunching mass at all the upper notches (200K and up) is not large enough to justify responses over the entire bracket above, whereas at the lower notches (100K-175K) this cannot be ruled out. We therefore focus on the upper notches in what follows. The other possibility is that non-bunching responses to notches affect the density throughout each bracket. This includes extensive responses as analyzed in detail in section 3.2. It could also include discrete intensive responses between the interiors of brackets, possibly driven by optimization

[83] Notches are located at round numbers and therefore provide an incentive to become a rounder by moving to the cutoff. For this reason, the non-rounder sample may understate behavioral responses as it captures bunching only by those who locate just below the cutoff and not by those who locate precisely at the cutoff.

frictions that prevent some individuals to target the region close to the notch point. Such effects cannot be ruled out and our bunching approach cannot capture them. Hence, while our approach fully accounts for frictions that prevent individuals from responding at all, it does not account for frictions that make people overshoot the notch point (beyond the narrow region of observed excess mass). If such effects are important, our estimates will be lower bounds.

The bottom panels of Figure 3.6 compare the empirical and counterfactual distributions around the four upper notches. The counterfactual (solid graph) is estimated for each notch separately by fitting a fifth-order polynomial to the empirical distribution, excluding data around the notch, as specified in (3.2.13).[84] The excluded range $[z_L, z_U]$ is demarcated by vertical dashed-black lines and the upper bound of the strictly dominated region is demarcated by a vertical dashed-red line.[85] Each panel shows estimates of excess bunching in proportion to the average counterfactual frequency in the dominated region (b), the share of individuals in the dominated region who are unresponsive (a^*), and the upper bound of the excluded range (z_U) ensuring that missing mass is equal to bunching mass.

The main findings are the following. First, excess bunching varies from 1.7 to 5.5 times the height of the counterfactual distribution across the different notches, and these estimates are strongly significant. Second, missing mass has a triangular shape and disappears to zero (point z_U) at about 35K-40K above each cutoff. This implies earnings responses of around 10% of income by the most elastic individuals. Third, despite the evidence of large bunching and missing mass, behavioral responses are strongly attenuated by optimization frictions: the share of individuals in dominated regions who are unresponsive is between 51% and 86% and precisely estimated. The amount of friction is negatively related to the amount of observed bunching across different notches. Fourth, since these notches create a discrete fall in consumption equal to 2.5% of gross income (with no change in leisure), our findings imply that a majority of the population face frictions (such as adjustment or attention costs) of at least 2.5% of gross income. Finally, using the approach developed in section 3.2, the amount of bunching absent frictions $b/\left(1-a^*\right)$ is 2 to 7 times larger than observed bunching b. Interestingly, the amount of bunching corrected for frictions is almost the same across different notches, suggesting that differences in observed bunching can be almost fully explained by differences in frictions.

Figure 3.7 turns to the full sample and is constructed exactly as the preceding figure.

[84] Figure A.12 in the appendix considers lower and higher polynomial degrees, showing that results are not very sensitive to this. Moreover, estimations are based on 500 rupee bins throughout and are not very sensitive to bin width.

[85] Note that the excluded range around some notches overlaps with the included range in the counterfactual estimation for other notches. Those overlaps do not have a big impact on the estimation, but more importantly they do not pose a conceptual problem: each notch is analyzed in isolation (as explained in section 3.2.3) and the locally estimated counterfactual is supposed to capture what would happen if the given notch were removed, taking all the other notches as given. For such an exercise, the bunching and missing mass regions at one notch should be seen as part of the counterfactual environment for other notches.

The empirical distribution for the full sample features larger excess mass at notch points than the distribution for non-rounders, but the full sample also features excess mass at other points that are not notches. The mass points between notches always occur at round numbers and their size depends on the roundness of the number in the annual and monthly dimension as described in section 3.2. There is more rounding at the bottom of the distribution than at the top, consistent with the earlier remark that rounding is a side-effect of poor record keeping. The counterfactual distribution is estimated as a fifth-order polynomial with round-number fixed effects as specified in (3.2.14). Estimates of excess bunching at notches (b) are net of round-number bunching in the counterfactual distribution. The findings for the full sample are qualitatively similar to those for the non-rounder sample: observed behavioral responses tend to be somewhat larger for the full sample while frictions are almost the same, implying that behavioral responses in the absence of frictions would be larger. Estimates for the full sample are generally not quite as robust to specification (e.g., polynomial degree) as estimates for the non-rounder sample.[86]

Taking advantage of the longitudinal aspect of the data, Table 3.1 investigates the dynamics and determinants of dominated and bunching behavior. We consider the non-rounder and full samples separately, distinguishing in each case between the unbalanced panel of those who file returns at least once during the sample period and the balanced panel of those who file returns every year. The table shows the total fractions featuring dominated and bunching behavior in each year as well as the fractions who have featured such behavior for two, three or four consecutive years. Bunchers include everybody locating in the bunching range $[z_L, z^*]$, only a subset of whom are *excess* bunchers actively responding to the tax system. The table explores misperception/inattention as a possible determinant of dominated behavior, using inconsistency between self-assessed tax liability and taxable income as an indicator of misperception (as described in section 3.3.1).[87]

The main insights are the following. First, the total fraction featuring dominated behavior declines over time and more so for the balanced sample of repeat filers who accumulate filing experience from year to year. Second, there is some persistence in dominated behavior from one year to the next, but almost everybody have moved out of such

[86] Figure A.13 in the appendix considers lower and higher polynomial degrees for the full sample.

[87] We assume that someone with taxable income in the dominated range *for income* is featuring dominated behavior even if his self-assessed tax liability is not in the corresponding dominated range *for tax payment*. This assumption relies on the efficacy of the automated validation system designed to flag and correct inconsistent returns (see also section 3.3.1). This system works as follows. Taxpayers who underestimate and underpay their income taxes (given their self-assessed taxable income) are labelled as "short filers" in Pakistan. The computer-based validation system generates a list of short filers along with automated notices asking those filers to remit the income tax they owe within two weeks. If short filers do not respond before the deadline, assessment orders are issued to recover the amount. Provided that this validation system is enforced without too much error (such that taxpayers do not find it optimal to *deliberately* display internal inconsistencies on their returns), our definition of dominated behavior is correct.

regions after three years. This shows the transitory nature of frictions at the individual level, but not necessarily at the aggregate level as new individuals come into dominated regions. Third, the total fraction featuring bunching behavior increases over time (more so for the balanced panel) and features stronger persistence over time than dominated behavior. Fourth, tax rate misperception is much more widespread among those in dominated regions (around 15-30%) than among those in bunching regions (around 5-10%), suggesting that misperception is a significant component of optimization frictions. Note that we should not expect zero misperception among bunchers, since we are considering everybody located in bunching regions and not just those who are actively responding to notches. Finally, Figure A.14 in the appendix shows graphically that excess bunching becomes stronger and dominated behavior slightly weaker over the sample period, consistent with the findings in Table 3.1. These findings together show that behavioral responses become less affected by frictions over time, so that the observed elasticity gets closer to the frictionless structural elasticity in the long run. With only four years of data, we cannot say if the long-run elasticity fully converges to the structural elasticity.

We now consider the estimation of structural elasticities, combining the non-parametric evidence above with the conceptual framework in section 3.2. Such elasticities can be obtained by estimating the earnings response of the marginal buncher and applying the parametric relationship (3.2.5) or the reduced-form approximation (3.2.12). We bound earnings responses and elasticities as described earlier: a lower bound is obtained from observed bunching scaled by the hole in the dominated region ("bunching-hole method" based on $b/(1-a^{*})$) and an upper bound is obtained from the point of convergence between the counterfactual and observed distributions ("convergence method" based on z_U). We focus on the non-rounder sample, because the inclusion of rounders has little impact on results while reducing precision. The results are presented in Table 3.2, which shows the notch point in column (1), the average tax rate jump in column (2), the size of the dominated range in column (3), frictions in the full dominated range and in the lower half of the dominated range in columns (4)-(5), earnings responses in columns (6)-(7), elasticities based on the parametric model in columns (8)-(9), and elasticities based on the reduced-form approximation in columns (10)-(11).[88]

The main findings are the following. First, the estimated amount of friction is almost the same in the lower part of the dominated region as in the full dominated region. This lends support to the assumption that frictions are locally constant, and it also suggests that the estimation is not biased by dynamic career effects as discussed in section 3.2.

[88] In the calculation of standard errors (using the bootstrap method described above), we impose the following constraints based on the theory. First, the earnings response is bottom-coded at the dominated range since this is the smallest possible response theoretically. Second, the earnings response based on the bunching-hole method is top-coded at the earnings response based on the convergence method as the latter represents an upper bound. Both of these constraints bind in less than 1% of the bootstrap iterations.

We therefore use the friction estimate based on the full dominated range in the rest of the table. Second, earnings responses are very large at all notches (5-15% of total earnings) and always precisely estimated. The magnitude of earnings responses reflects the combination of large observed bunching and large frictions. Third, the structural elasticities driving those large earnings responses are in general modest except at 200K. The lower-bound elasticities fall mostly in the interval 0.05-0.15 (0.30 at 200K) while the upper-bound elasticities fall mostly in the interval 0.10-0.25 (above 1 at 200K). The combined findings of large bunching responses and small structural elasticities highlights the mechanism design problem with notches. Fourth, elasticities are not as precisely estimated as earnings responses because of the strong nonlinearity of the formula that links the elasticity to the earnings response. While elasticities based on the bunching-hole method are almost always statistically significant, elasticities based on the convergence method are often not significant. Finally, note that estimates of observed elasticities attenuated by frictions can be obtained by multiplying the elasticities in the table by the share of responders $1 - a^*$. This exercise implies observed elasticities extremely close to zero.

The smallness of elasticity estimates may be surprising as these are structural (frictionless) elasticities of taxable income (real and evasion responses) among self-employed individuals in a context of weak enforcement. The following points are worth keeping in mind when thinking about the small magnitudes. First, the population of tax filers in Pakistan is likely to be a selected sample of individuals who are relatively well-monitored and/or have high tax morale, dampening the evasion channel of taxable income response to tax rates. Second, even if enforcement is weak and evasion therefore large, this does not necessarily imply a large evasion *response* to tax rate *changes*. Theoretically, the evasion response to tax rate changes depends on the curvature—not the level—of detection probabilities and penalties as a function of evasion (Kleven *et al.* 2011), and even the sign of the effect is in general ambiguous. Empirically, we are not aware of any previous study showing compelling evidence of large evasion responses to tax rates even in samples featuring large evasion levels (Kleven *et al.* 2011; Saez *et al.* 2012). Third, since our approach does not capture extensive responses (including informality) and potential discrete intensive responses between the interiors of brackets, we cannot conclude that the *total* elasticity of taxable income is necessarily small in Pakistan.

3.3.4 Results for Wage Earners

For wage earners, we focus on the non-rounder sample throughout. Rounding is much less of an issue for wage earners (only 8 percent report income in even thousands) than for self-employed individuals as they have access to more accurate income records. Given the small share of rounders, including them has no substantive effect on conclusions.

The tax schedule for wage earners has twenty notches, but we concentrate on six notches in the middle of the schedule (400K-950K). The bottom notches offer less compelling variation because they are small and occur in a range where many wage earners are affected by filing exemptions. The top notches are not very useful because they occur in the extreme tail of the distribution where the density distribution is too noisy for a precise bunching analysis. Figure 3.8 presents non-parametric evidence on behavioral responses and frictions for wage earners in 2006-2007, and is constructed exactly as the analogous figures in the previous section.[89] The estimation of the counterfactual distribution is based on a third-order polynomial (instead of a fifth-order polynomial above) as the distribution for wage earners has less curvature than the distribution for self-employed individuals.

The main findings in Figure 3.8 are the following. First, the empirical distribution features sharp bunching below every notch along with clear missing mass above every notch. Unlike the findings for self-employed individuals, missing mass appears as a clear hole above several of the notches. Second, the empirical distribution does not feature the step-function pattern observed for self-employed individuals, but a missing mass area that is flat or increasing after which the density is smoothly declining until the next notch. This is consistent with the conceptual Figure A.9 (Panel C) in the appendix. It suggests that the possible explanations for the step-pattern—large bunching responses over multiple notches or non-bunching responses affecting the entire bracket above the cutoff—are not present for wage earners. Third, unsurprisingly bunching is not as large for wage earners as it is for self-employed individuals, although it should be noted that the notches for wage earners are considerably smaller. Excess bunching is between 30% and 80% of the height of the counterfactual frequency and precisely estimated. Fourth, frictions are considerably larger for wage earners than for self-employed individuals, with as many as 90% of wage earners in strictly dominated ranges being unresponsive to notches. This provides direct evidence that adjustment costs in earnings severely constrain behavioral responses for wage earners, who are often bound by fixed wage-hours contracts in the short run. Our findings imply that, if not for such constraints, bunching for wage earners would be 10 times larger that what we observe.

Table 3.3 presents estimates of earnings responses and structural elasticities, and is constructed like the corresponding table in the previous section. The key findings are the following. First, the estimation of frictions is virtually unchanged as we zoom in on the bottom half of the dominated range, lending further support to the bunching-hole method based on the assumption of locally constant frictions. Second, earnings responses are mostly between 2 and 5 percent of earnings across the different notches. Those responses are precisely estimated when using the bunching-hole method, but not the convergence method. Third, those earnings responses are driven by very small structural elasticities,

[89] Unlike the previous section, we show evidence for males and females together as the middle notches we focus on apply to both groups.

generally around 0.05 or lower.

3.3.5 Shifting Between Self-Employment and Wage Income

We now analyze the income-composition notch described earlier: each individual is classified as self-employed (wage earner) if the share of self-employment income in total income is greater than or equal to (smaller than) 50%, with much higher tax rates on the self-employed than on wage earners. This creates a large notch at a self-employment income share of 50% and provides very strong incentives to change the composition of income (e.g. through income shifting) to obtain the more lenient tax treatment.

Figure 3.9 presents non-parametric evidence of behavioral responses to this income-composition notch. Panel A shows the empirical distribution of the self-employment income share as a histogram in 1% bins. We exclude the end points of 0% (only wages) or 100% (only self-employment income), which accounts for most of the population and feature huge mass points. Unlike the notches considered earlier, the cutoff itself belongs to the high-tax region and we therefore expect to see bunching only strictly below the notch. To evaluate this, each bin excludes the upper bound of the interval such that the notch point belongs to the bin above rather than below. The following findings emerge from the figure. First, there is a clear behavioral response as the distribution features large excess mass on the low-tax side and large missing mass on the high-tax side of the notch. Second, bunching is more diffuse than seen earlier, and it is not possible to explain missing mass above the notch by bunching mass in a narrow range below the notch. This suggests that some individuals respond by substantially overshooting the cutoff. Third, surprisingly there is excess bunching in the first bin above the notch. It turns out that all of this excess bunching is driven by individuals with a self-employment income share *exactly* equal to 50%, which points to two possible explanations: (i) it can be a form of "round-number bunching" by individuals who do not know their income composition and therefore report the same amount of self-employment and wage income, or (ii) it can be a bunching response by individuals who do not know that the cutoff itself (unlike all other notches in the tax system) belongs to the high-tax range. In the first case it is natural to drop the round-number observations at 50%, while in the second case it is natural to relocate those observations to the bin below. In Panel B, we take the more conservative option of dropping observations at the cutoff.

Panel B compares the empirical distribution to a counterfactual distribution, estimated by fitting a fifth-order polynomial to the observed bin counts excluding observations in a range $[s_L, s_U]$. To account for the fact that missing mass cannot be justified by excess mass in a narrow range below the notch, the lower bound s_L must be located farther into the interior of the lower bracket. The lower bound is determined visually at a point where the declining distribution appears to flatten and feature a kink. The upper bound

s_U is estimated to ensure that missing mass in the range $[0.50, s_U]$ equals excess mass in the range $[s_L, 0.50)$. We find the following. First, excess bunching equals 11.4 times the average height of the counterfactual distribution in the bunching range, but is not precisely estimated. Second, the upper bound of the excluded range equals 87% and is precisely estimated, implying that the most responsive individuals reduce their self-employment income share by 37%-points (or possibly more given that some them overshoot the notch point). Finally, while these are extremely large behavioral responses, the size of the notch is also truly massive: the tax rate jump between wage-earner and self-employment status (for a given level of total income) is on average 6 percentage points among the filers in Figure 3.9, considerably larger than the notches considered earlier.[90]

3.4 Conclusion

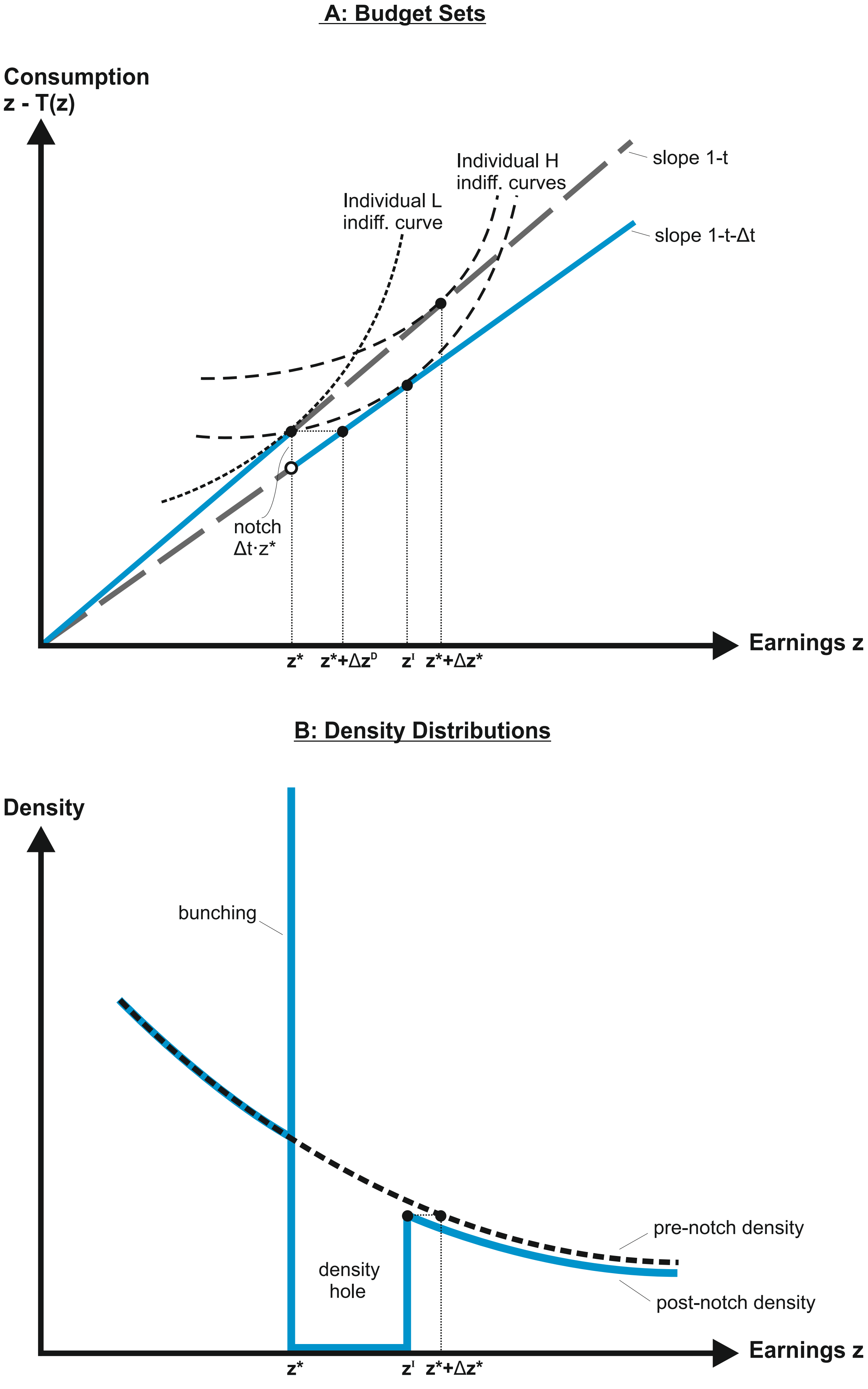
Notches are widespread in tax and transfer systems around the world, but have not been systematically explored in empirical work. We show that notches often create regions of strictly dominated choice that would be empty in the absence of optimization frictions, implying that observed density mass in such regions non-parametrically identifies frictions. By combining estimates of frictions and excess bunching at notches, it is possible to separately identify the observed elasticity attenuated by frictions and the structural elasticity absent frictions. If frictions disappear over long time horizons, the structural elasticity represents a long-run parameter that determines welfare and optimal policy. Using longitudinal data, notches can be used to analyze how frictions evolve over time and whether the observed elasticity does in fact converge to the structural elasticity in the long run. The conceptual approach developed here represents a significant advance over existing approaches based on kinks and tax reforms, which cannot shed light on frictions and true structural parameters without strong parametric assumptions.

Applying our framework to tax notches in Pakistan, we demonstrate the power of the approach and present the first compelling evidence of behavioral responses to taxes in a developing country. The most striking finding is perhaps the quantitative importance of frictions: despite the extremely strong tax incentives created by notches, the majority of the population are unresponsive to those incentives. This contradicts the conventional view that behavioral responses to large tax changes are not attenuated by frictions and therefore represent long-run effects. Another striking finding is that, absent attenuation bias from frictions, behavioral responses to notches are very large while the structural

[90] It is conceptually difficult to turn these estimates into a structural elasticity without knowing the anatomy of the composition response. In particular, the structural elasticity will depend on whether this is a pure *shifting* response (changing composition for a given level of total income) or if it is partly or fully a *level* response (such as reducing self-employment income for a given level of wages). Our data do not permit us to clearly identify the mechanism driving composition bunching due to lack of power and the diffuseness of bunching.

elasticities driving those responses are in general modest. This highlights the efficiency problem with notches: by creating extremely large implicit marginal tax rates around cutoffs, they induce very large behavioral responses and efficiency costs even when structural elasticities are small.

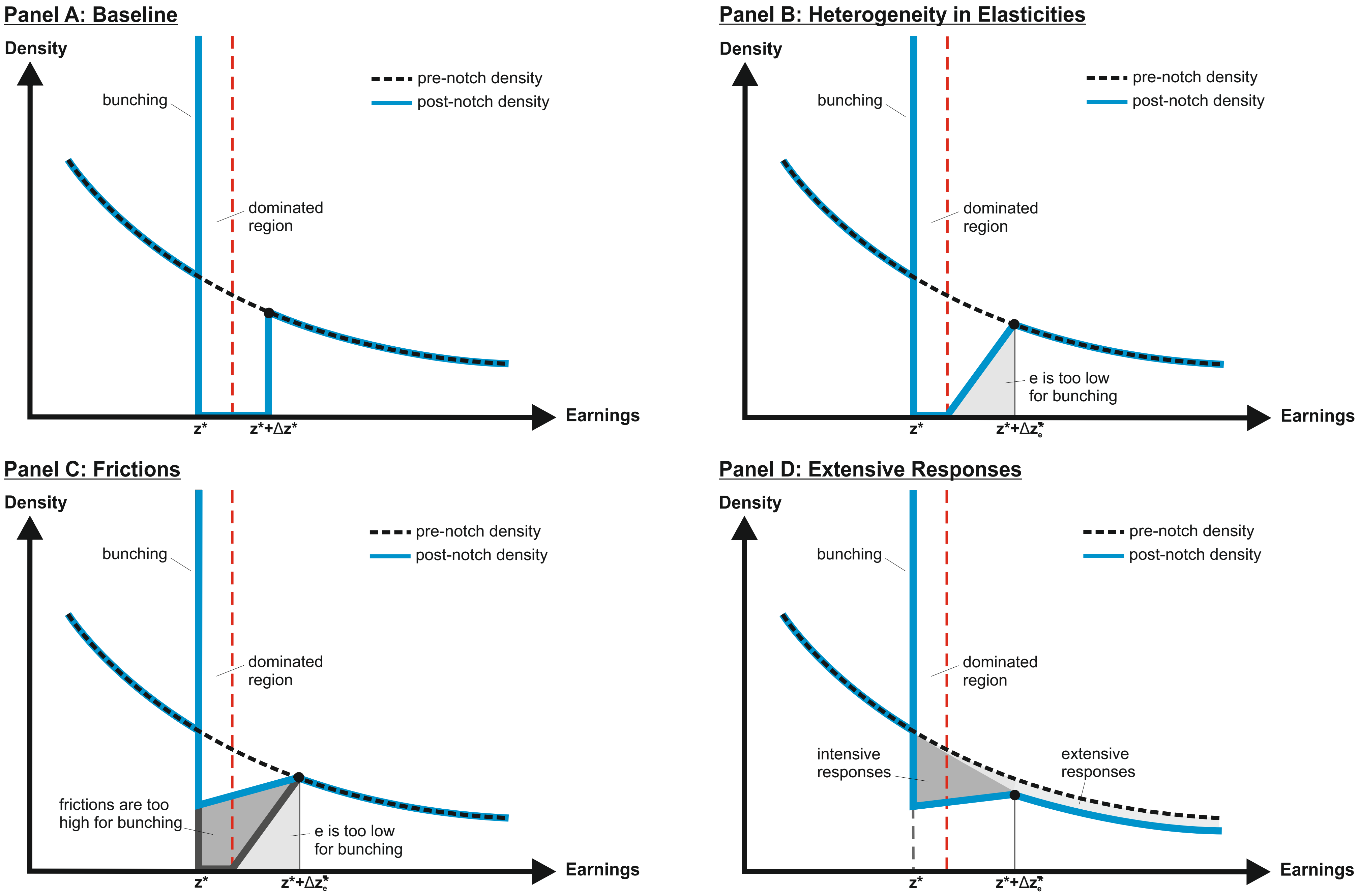
Figure 3.1
Behavioral Responses to a Tax Notch



A: Budget Sets

B: Density Distributions

Figure 3.2
Density Distributions Under Different Model Extensions



Panel A: Baseline

Panel B: Heterogeneity in Elasticities

Panel C: Frictions

Panel D: Extensive Responses

Figure 3.3

Reduced-Form Approximation of Earnings Elasticity

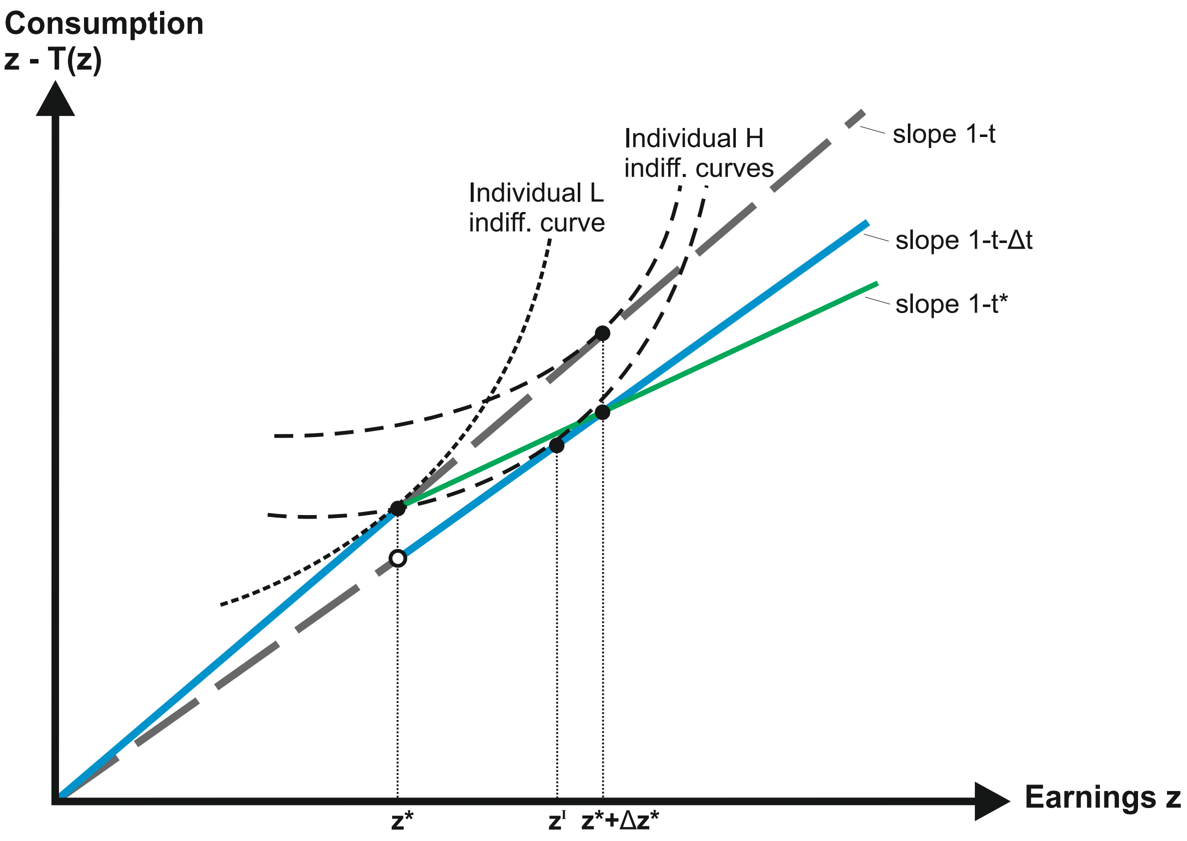
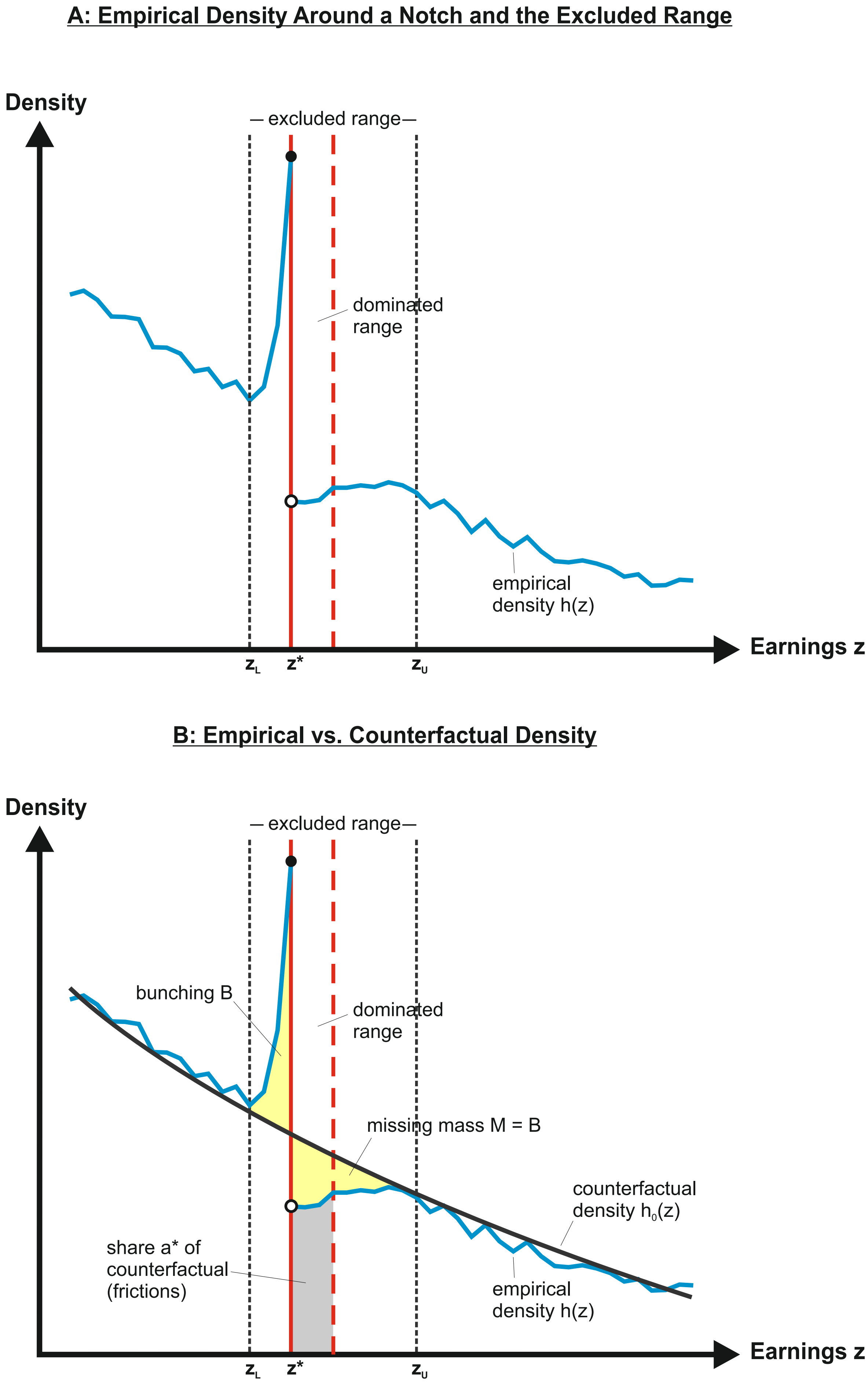


Figure 3.4
Estimating the Counterfactual Density from an Empirical Density

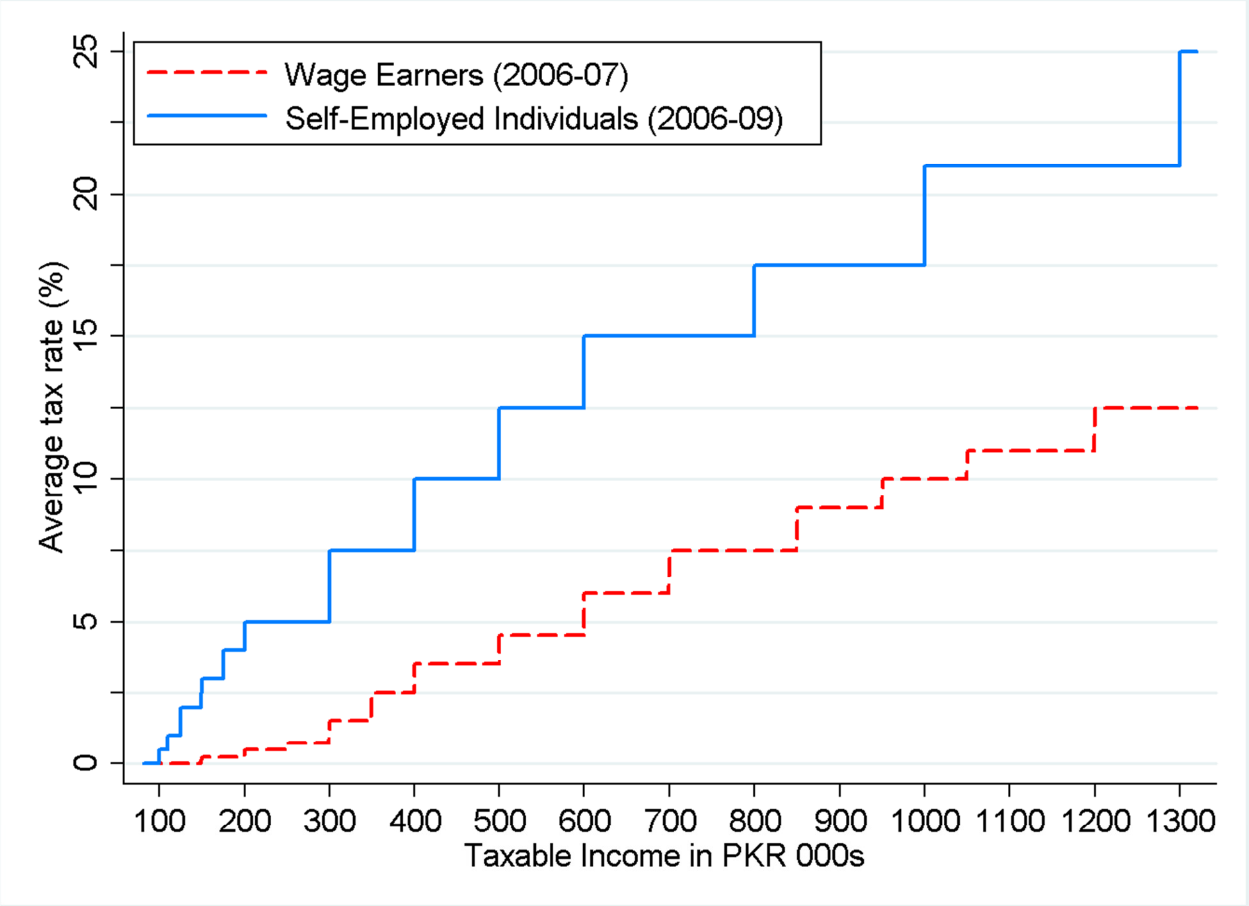


A: Empirical Density Around a Notch and the Excluded Range

B: Empirical vs. Counterfactual Density

Figure 3.5

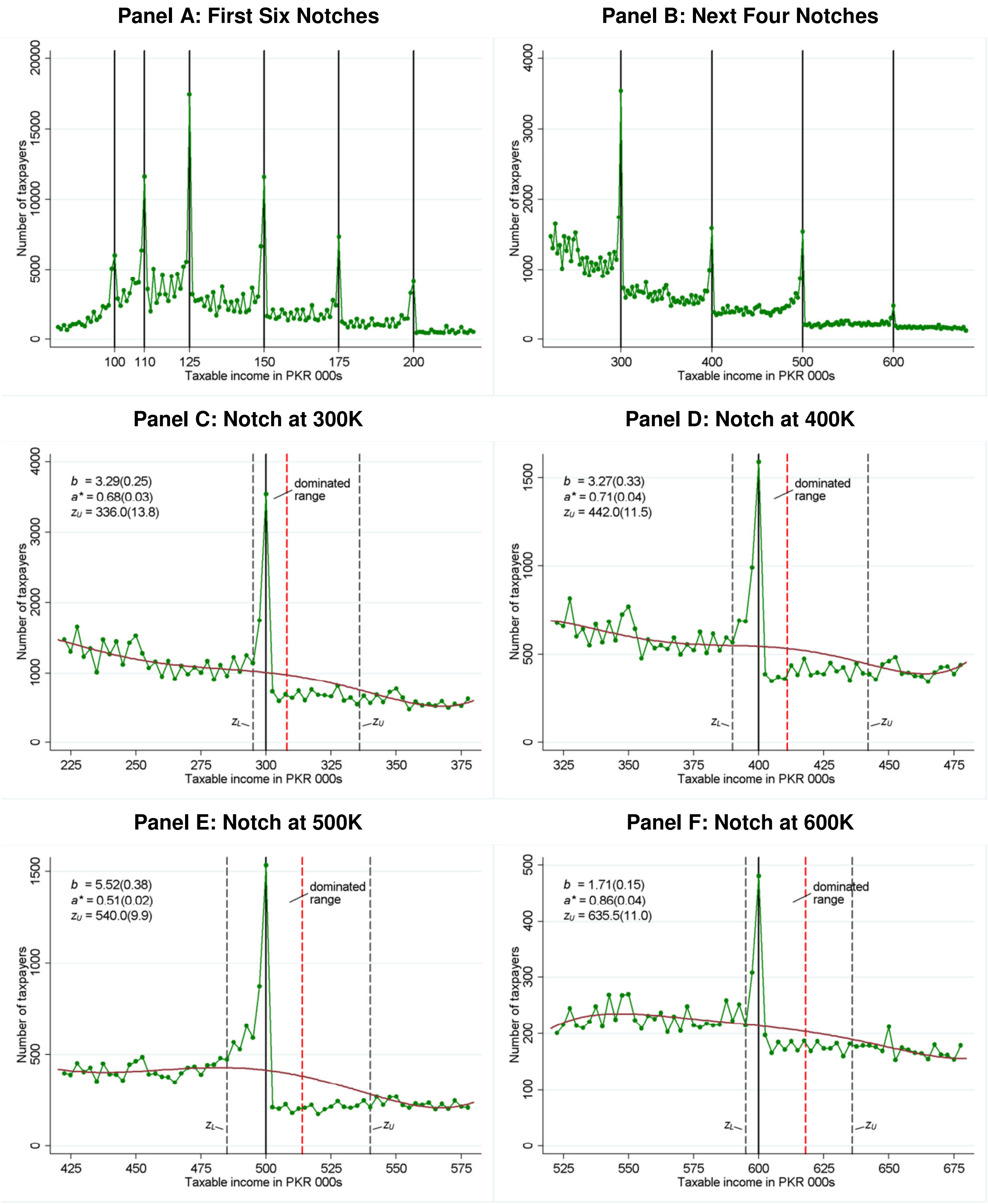
Personal Income Tax Schedules in Pakistan



Notes: the figure shows the average tax rate as a function of annual taxable income for wage earners in 2006-07 (red dashed line) and for self-employed individuals in 2006-09 (blue solid line). Taxable income is shown in thousands of Pakistani Rupees (PKR), with the PKR-USD exchange rate varying from 60-80 during these years. Each bracket cutoff is associated with a discrete jump in the average tax rate (a notch), and the cutoff itself belongs to the low-tax side of the notch. The tax rate on self-employed individuals increases from 0 to 25% over thirteen notches, while the tax rate on wage earners increases from 0 to 20% over twenty notches (the first thirteen of which are shown in the figure). The tax system classifies an individual as self-employed (wage earner) if self-employment income as a share of total income is greater than or equal to (less than) 50%, and then taxes total income according to the assigned schedule.

Figure 3.6

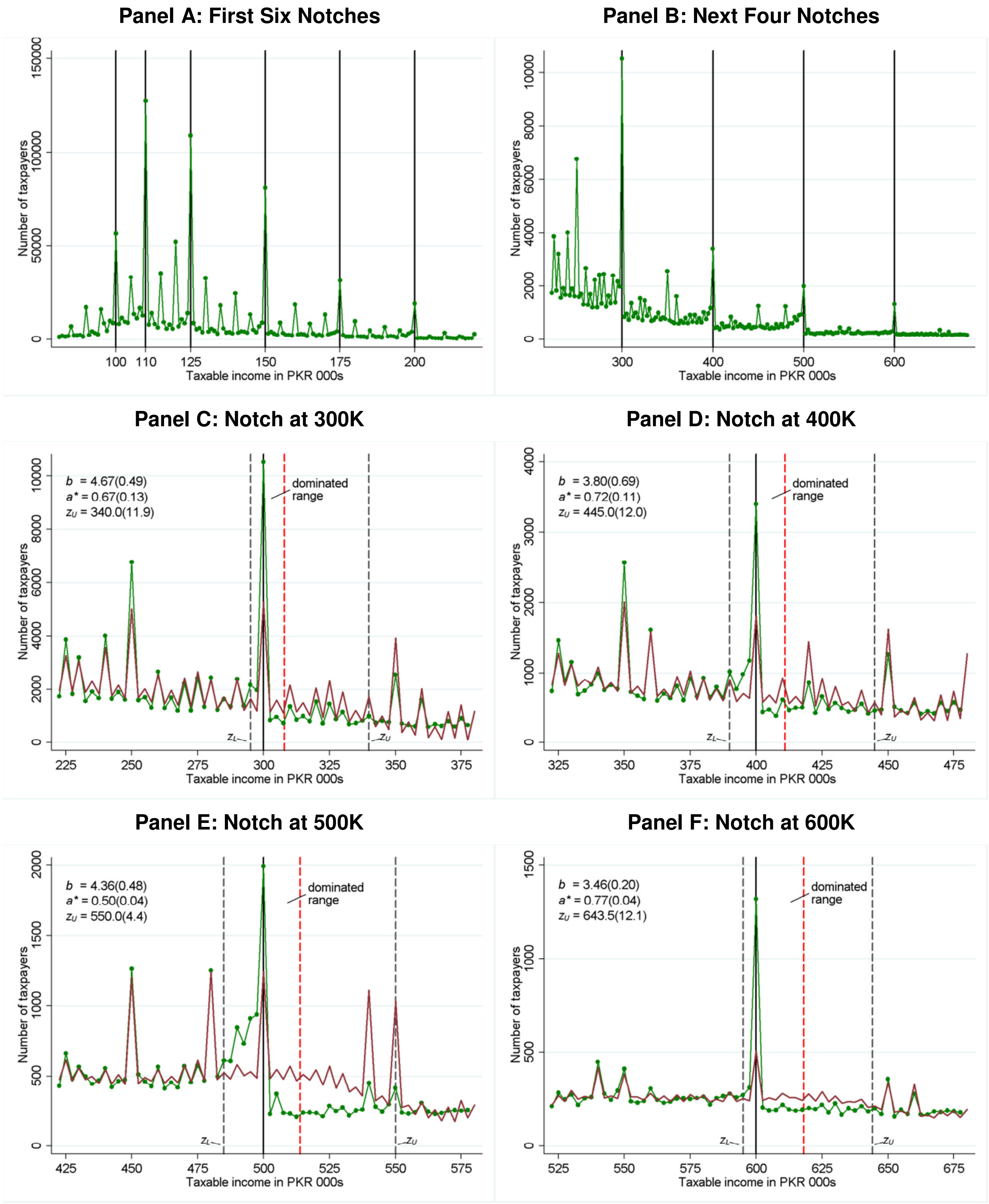
Empirical and Counterfactual Distributions around Notches:
Self-Employed Individuals (Non-Rounder Sample)



Notes: the figure shows the empirical distribution of taxable income (dotted green graph) and the counterfactual distribution (solid brown graph) for self-employed individuals (non-rounder sample) from 2006-09. The counterfactual is estimated for each notch separately by fitting a fifth-order polynomial to the empirical distribution, excluding data around the notch, as specified in equation (13). Notch points are marked by solid black lines, upper bounds of dominated regions are marked by dashed red lines, and excluded ranges [z_L,z_U] are marked by dashed black lines. Bunching b is excess mass in the excluded range below the notch (in proportion to the average counterfactual frequency in the dominated range), a^* is the share of individuals in the dominated range who are unresponsive, and the upper bound of the excluded range z_U has been estimated to ensure that missing mass equals bunching mass. Standard errors are shown in parentheses.

Figure 3.7

Empirical and Counterfactual Distributions around Notches:
Self-Employed Individuals (Full Sample)



Panel A: First Six Notches

Panel B: Next Four Notches

Panel C: Notch at 300K

Panel D: Notch at 400K

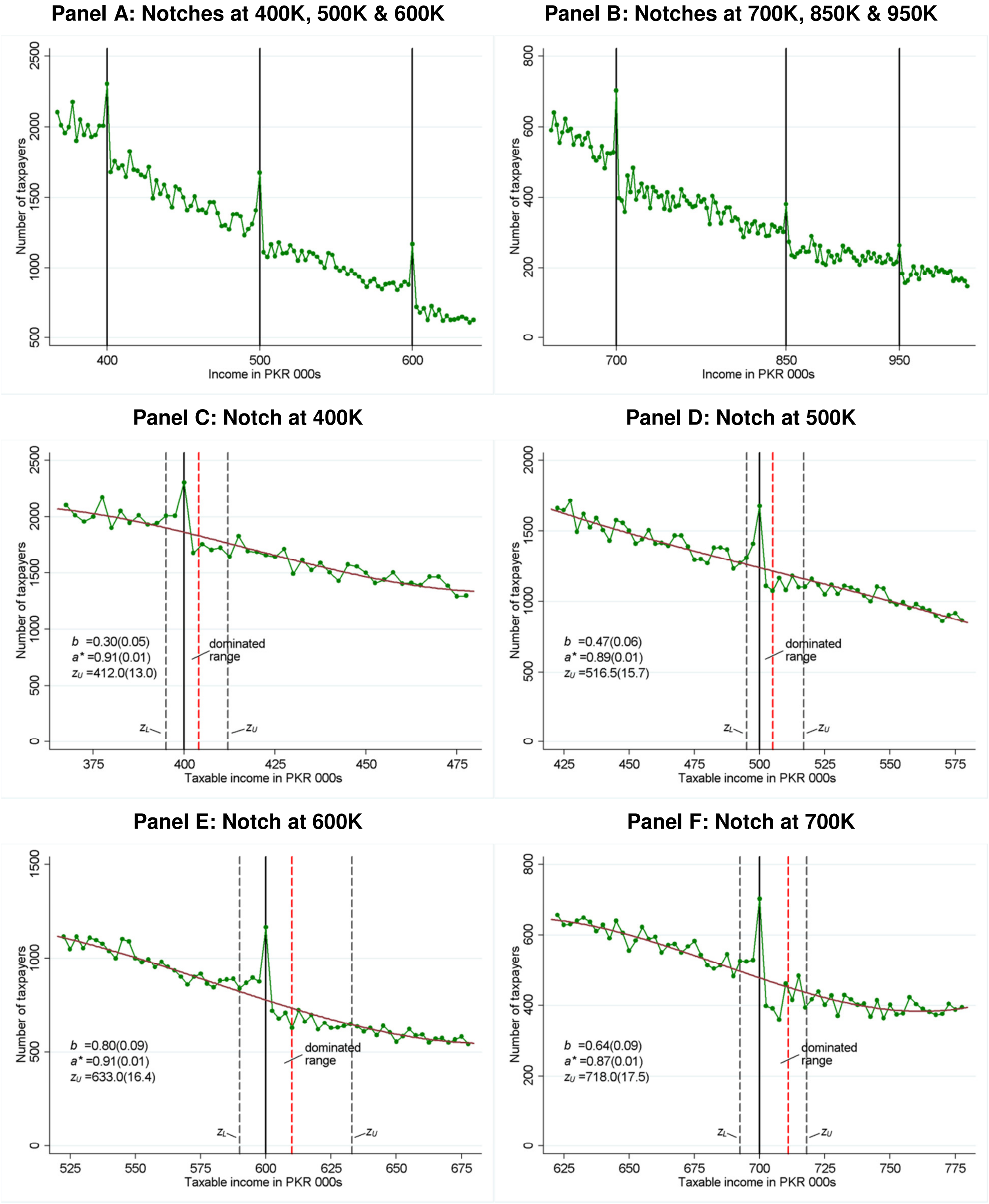
Panel E: Notch at 500K

Panel F: Notch at 600K

Notes: the figure shows the empirical distribution of taxable income (dotted green graph) and the counterfactual distribution (solid brown graph) for self-employed individuals (full sample) from 2006-09. The counterfactual is estimated for each notch separately by fitting a fifth-order polynomial with round-number fixed effects to the empirical distribution, excluding data around the notch, as specified in equation (14). Notch points are marked by solid black lines, upper bounds of dominated regions are marked by dashed red lines, and excluded ranges $[z_L,z_U]$ are marked by dashed black lines. Bunching *b* is excess mass in the excluded range below the notch (in proportion to the average counterfactual frequency in the dominated range), *a** is the share of individuals in the dominated range who are unresponsive, and the upper bound of the excluded range z_U has been estimated to ensure that missing mass equals bunching mass. Standard errors are shown in parentheses.

Figure 3.8

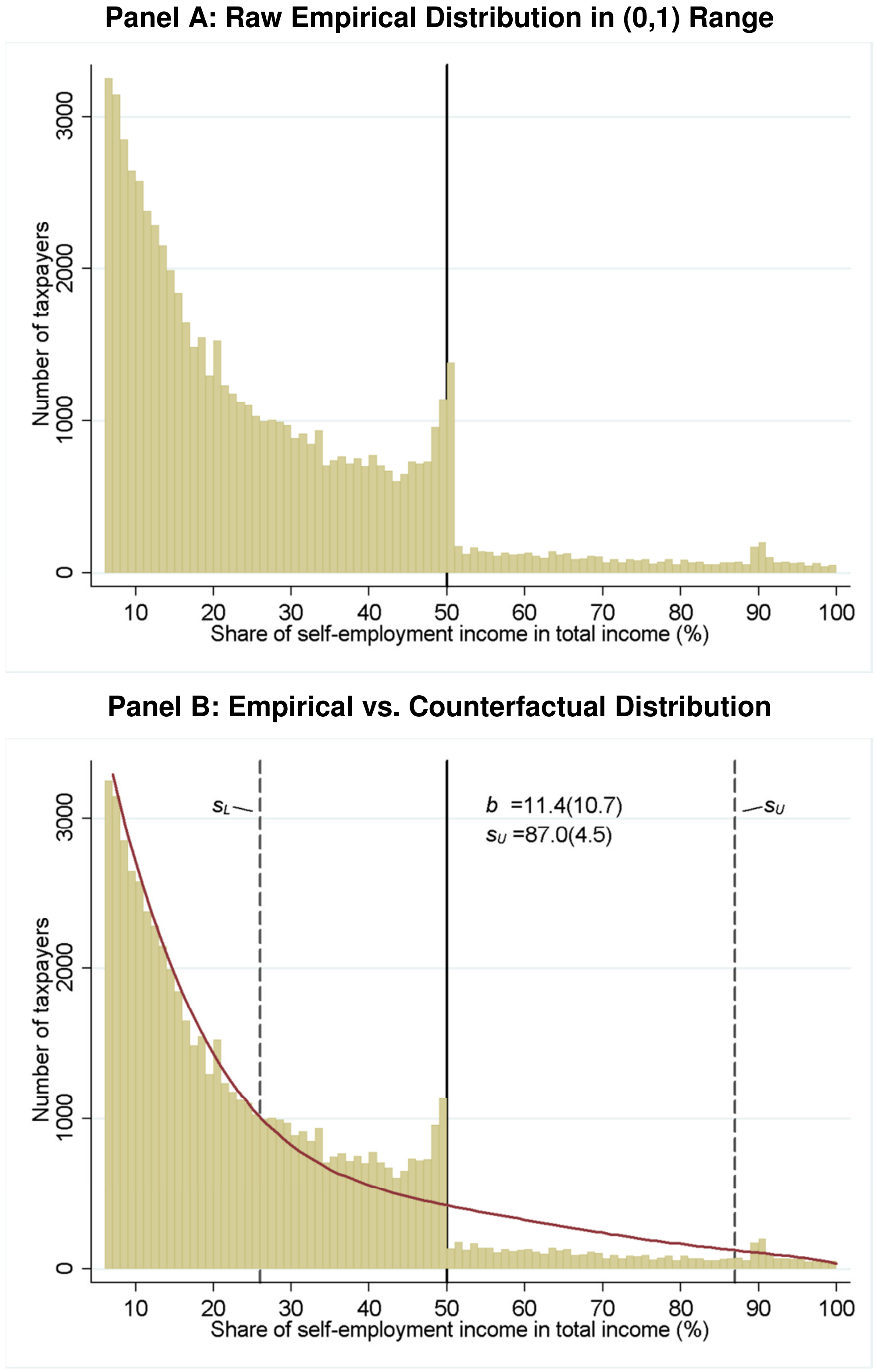
Empirical and Counterfactual Distributions around Notches:
Wage Earners (Non-Rounder Sample)



Notes: the figure shows the empirical distribution of taxable income (dotted green graph) and the counterfactual distribution (solid brown graph) for wage earners (non-rounder sample) from 2006-07. The counterfactual is estimated for each notch separately by fitting a third-order polynomial to the empirical distribution, excluding data around the notch, as specified in equation (13). Notch points are marked by solid black lines, upper bounds of dominated regions are marked by dashed red lines, and excluded ranges [z_L,z_U] are marked by dashed black lines. Bunching *b* is excess mass in the excluded range below the notch (in proportion to the average counterfactual frequency in the dominated range), *a** is the share of individuals in the dominated range who are unresponsive, and the upper bound of the excluded range z_U has been estimated to ensure that missing mass equals bunching mass. Standard errors are shown in parentheses.

Figure 3.9

Empirical and Counterfactual Distributions of the Self-Employment Income Share:
Behavioral Responses to the Income-Composition Notch at 50%



Panel A: Raw Empirical Distribution in (0,1) Range

Panel B: Empirical vs. Counterfactual Distribution

Notes: the figure shows the empirical distribution of the self-employment income share (in bars) and the counterfactual distribution (solid brown graph) for individuals with both self-employment income and wage income from 2006-09. Panel A includes all observations with a self-employment income share between zero and one, while Panel B drops observations precisely at the cutoff value of 50%. The counterfactual is estimated by fitting a fifth-order polynomial to the empirical distribution, excluding data around the notch, as specified in equation (13). The Notch point is marked by a solid black line and the excluded range [s_L,s_U] is marked by dashed black lines. Bunching *b* is excess mass in the excluded range below the notch (in proportion to the average counterfactual frequency in this range), and the upper bound of the excluded range s_U has been estimated to ensure that missing mass equals bunching mass. Standard errors are shown in parentheses.

Table 3.1

Dynamics of Dominated, Bunching and Inconsistent Behavior

Year	#Obs.	Dominated Behavior				Bunching Behavior				Inconsistent Reporting		
		Total	2-Year	3-Year	4-Year	Total	2-Year	3-Year	4-Year	Total	Among Dominated	Among Bunchers
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Panel A: Non-Rounder Sample												
A1: Unbalanced Panel												
2006	130,037	**5.1**	-	-	-	**23.3**	-	-	-	**12.2**	**17.1**	**9.4**
		(0.06)				(0.12)				(0.09)	(0.46)	(0.17)
2007	129,443	**4.4**	**0.43**	-	-	**25.2**	**6.1**	-	-	**10.6**	**15.0**	**8.1**
		(0.06)	(0.018)			(0.12)	(0.07)			(0.09)	(0.47)	(0.15)
2008	130,110	**4.4**	**0.36**	**0.09**	-	**25.9**	**7.3**	**2.2**	-	**10.2**	**13.1**	**8.0**
		(0.06)	(0.017)	(0.008)		(0.12)	(0.07)	(0.04)		(0.08)	(0.44)	(0.15)
2009	108,331	**4.6**	**0.30**	**0.05**	**0.03**	**27.7**	**6.9**	**2.5**	**0.9**	**15.3**	**30.1**	**5.1**
		(0.06)	(0.017)	(0.007)	(0.005)	(0.14)	(0.08)	(0.05)	(0.03)	(0.11)	(0.65)	(0.13)
A2: Balanced Panel												
2006	30,120	**4.3**	-	-	-	**25.2**	-	-	-	**11.6**	**14.7**	**9.3**
		(0.12)				(0.25)				(0.18)	(0.99)	(0.33)
2007	30,120	**3.6**	**0.47**	-	-	**27.2**	**10.7**	-	-	**11.7**	**16.1**	**9.0**
		(0.11)	(0.039)			(0.26)	(0.18)			(0.19)	(1.11)	(0.32)
2008	30,120	**3.8**	**0.37**	**0.18**	-	**27.4**	**11.9**	**5.2**	-	**11.7**	**16.9**	**9.6**
		(0.11)	(0.035)	(0.024)		(0.26)	(0.19)	(0.13)		(0.19)	(1.11)	(0.32)
2009	30,120	**3.6**	**0.41**	**0.12**	**0.11**	**32.0**	**13.5**	**6.5**	**3.1**	**7.1**	**16.1**	**3.0**
		(0.11)	(0.037)	(0.020)	(0.019)	(0.27)	(0.20)	(0.14)	(0.10)	(0.15)	(1.12)	(0.17)
Panel B: Full Sample												
B1: Unbalanced Panel												
2006	405,260	**2.5**	-	-	-	**33.5**	-	-	-	**8.4**	**15.1**	**7.0**
		(0.02)				(0.07)				(0.04)	(0.36)	(0.07)
2007	393,925	**2.1**	**0.22**	-	-	**35.2**	**13.6**	-	-	**7.8**	**14.1**	**6.3**
		(0.02)	(0.007)			(0.08)	(0.05)			(0.04)	(0.38)	(0.06)
2008	379,962	**2.2**	**0.18**	**0.04**	-	**36.0**	**15.6**	**7.1**	-	**8.2**	**12.9**	**6.4**
		(0.02)	(0.007)	(0.003)		(0.08)	(0.06)	(0.04)		(0.04)	(0.37)	(0.07)
2009	324,901	**2.2**	**0.15**	**0.02**	**0.01**	**40.7**	**15.4**	**7.9**	**4.0**	**8.5**	**26.0**	**3.3**
		(0.03)	(0.007)	(0.003)	(0.002)	(0.09)	(0.06)	(0.05)	(0.03)	(0.05)	(0.51)	(0.05)
B2: Balanced Panel												
2006	167,500	**2.5**	-	-	-	**35.0**	-	-	-	**8.2**	**13.2**	**6.9**
		(0.04)				(0.12)				(0.07)	(0.52)	(0.10)
2007	167,500	**1.9**	**0.26**	-	-	**36.7**	**17.7**	-	-	**8.0**	**13.8**	**6.5**
		(0.03)	(0.012)			(0.12)	(0.09)			(0.07)	(0.61)	(0.10)
2008	167,500	**1.9**	**0.18**	**0.06**	-	**37.3**	**19.4**	**10.8**	-	**8.7**	**14.4**	**6.8**
		(0.03)	(0.010)	(0.006)		(0.12)	(0.10)	(0.08)		(0.07)	(0.62)	(0.10)
2009	167,500	**1.8**	**0.19**	**0.04**	**0.02**	**43.0**	**21.7**	**12.8**	**7.7**	**3.7**	**11.4**	**1.9**
		(0.03)	(0.011)	(0.005)	(0.004)	(0.12)	(0.10)	(0.08)	(0.06)	(0.05)	(0.57)	(0.05)

Notes: the table shows the dynamics of dominated, bunching, and inconsistent behavior for self-employed individuals from 2006-2009. Panel A and B show the non-rounder and full samples, respectively, and each panel distinguishes between the unbalanced panel (everybody filing at least once in the sample period) and the balanced panel (those filing every year in the sample period). The table shows the total fractions featuring dominated or bunching behavior in each year as well as the fractions featuring such behavior for two, three or four consecutive years. Bunchers include everybody locating in the bunching range below one of the thirteen notches, only a subset of whom are excess bunchers actively responding to the tax system. The table considers misperception as a determinant of dominated behavior, using inconsistency between self-assessed tax liability and taxable income as an indicator.

Table 3.2

Structural Earnings Elasticities for Self-Employed Individuals (Non-Rounder Sample)

Notch Point	ATR Jump Δt	Dominated Range Δz^D	Frictions a*		Earnings Response Δz^*		Structural Elasticity e		Reduced-Form Elasticity e_R	
			Full Range Δz^D	Lower Range $\Delta z^D/2$	Bunching-Hole Method	Conv. Method	Bunching-Hole Method	Conv. Method	Bunching-Hole Method	Conv. Method
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
200K	1.0	2,105	0.531***	0.503***	17,000***	34,000***	0.281*	1.021**	0.333*	1.279**
			(0.036)	(0.033)	(4,268)	(7,337)	(0.159)	(0.430)	(0.188)	(0.636)
300K	2.5	8,108	0.682***	0.683***	27,500***	36,000**	0.107***	0.188	0.153***	0.258
			(0.029)	(0.028)	(4,492)	(13,766)	(0.039)	(0.188)	(0.050)	(0.266)
400K	2.5	11,111	0.712***	0.684***	29,500***	42,000***	0.065	0.171*	0.097**	0.194
			(0.037)	(0.033)	(6,760)	(11,451)	(0.044)	(0.095)	(0.046)	(0.131)
500K	2.5	14,286	0.512***	0.532***	30,500***	40,000***	0.062***	0.079	0.065***	0.111
			(0.021)	(0.020)	(3,908)	(9,876)	(0.014)	(0.053)	(0.017)	(0.074)
600K	2.5	17,647	0.861***	0.849***	31,500***	35,500***	0.025**	0.035	0.047***	0.060
			(0.035)	(0.031)	(5,742)	(10,952)	(0.012)	(0.036)	(0.015)	(0.047)

Notes: considering the sample of self-employed individuals (non-rounders) from 2006-2009, the table presents estimates of frictions (share of individuals in dominated ranges who are unresponsive) in columns (4)-(5), earnings responses to notches absent frictions in columns (6)-(7), and structural earnings elasticities based on either the parametric equation (5) in columns (8)-(9) or the reduced-form approximation (12) in columns (10)-(11). The bunching-hole method scales observed bunching B by the inverse of the hole in the dominated range 1/(1-a*) in order to estimate responses that are not attenuated by optimization frictions. The convergence method estimates responses based on the point of convergence between the observed and counterfactual densities. Those two methods provide lower and upper bounds on the average structural elasticity in the population. Standard errors are shown in parentheses and stars indicate statistical significance level (* = 10% level, ** = 5% level, *** = 1% level).

Table 3.3

Structural Earnings Elasticities for Wage Earners (Non-Rounder Sample)

Notch Point	ATR Jump Δt	Dominated Range Δz^D	Frictions a*		Earnings Response Δz^*		Structural Elasticity e		Reduced-Form Elasticity e_R	
			Full Range Δz^D	Lower Range $\Delta z^D/2$	Bunching-Hole Method	Conv. Method	Bunching-Hole Method	Conv. Method	Bunching-Hole Method	Conv. Method
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
400K	1.0	4,145	0.914***	0.908***	9,000***	12,000	0.024***	0.031	0.024**	0.043
			(0.013)	(0.013)	(1,984)	(12,969)	(0.009)	(0.184)	(0.011)	(0.218)
500K	1.0	5,236	0.890***	0.899***	11,000***	16,500	0.035***	0.038	0.023**	0.052
			(0.008)	(0.008)	(2,045)	(15,702)	(0.010)	(0.169)	(0.009)	(0.209)
600K	1.0	9,574	0.913***	0.908***	24,000***	33,000**	0.034*	0.074	0.050**	0.094
			(0.011)	(0.011)	(5,310)	(16,381)	(0.019)	(0.113)	(0.023)	(0.132)
700K	1.5	11,351	0.870***	0.834***	12,500***	18,000	0.001	0.010	0.010***	0.020
			(0.010)	(0.010)	(1,928)	(17,476)	(0.013)	(0.064)	(0.003)	(0.079)

Notes: considering the sample of wage earners (non-rounders) from 2006-2007, the table presents estimates of frictions (share of individuals in dominated ranges who are unresponsive) in columns (4)-(5), earnings responses to notches absent frictions in columns (6)-(7), and structural earnings elasticities based on either the parametric equation (5) in columns (8)-(9) or the reduced-form approximation (12) in columns (10)-(11). The bunching-hole method scales observed bunching B by the inverse of the hole in the dominated range 1/(1-a*) in order to estimate responses that are not attenuated by optimization frictions. The convergence method estimates responses based on the point of convergence between the observed and counterfactual densities. Those two methods provide lower and upper bounds on the average structural elasticity in the population. Standard errors are shown in parentheses and stars indicate statistical significance level (* = 10% level, ** = 5% level, *** = 1% level).

Appendix

A.1 Details of Firm Characteristics

1. **Large.** I divide the firms on the basis of their average sales in three pre-reform years, 2006-08. The dummy variable takes the value 1 for firms with sales above the 75th percentile of the size distribution.[91]

2. **Electronic Return Filer.** All partnership firms were required to file returns electronically for the years 2008-11. Some of the firms did not comply with this mandatory provision, while a few were filing electronically even before the provision came into effect. I categorize a firm electronic filer if any of the four returns for years 2006-09 was filed electronically (about 80% of the firms). The variable, thus, captures minimum level of sophistication needed to be a functional tax payer.

3. **Registered for VAT.** The variable indicates if the firms was registered with the FBR to remit VAT on its sales (about 30% of the firms in balanced panel sample).

4. **Age.** The dummy variable takes the value 1 if age of the firm – measured in number of years since the firm registered with the FBR – was more than the 75th percentile (6 years).

5. **Withholding Agent.** The variable is an indicator for the firms that acted as withholding agents, collecting tax on behalf of the government (about 22% of the firms).

6. **Third Party Reporting.** Pakistani tax code stipulates a comprehensive tax withholding scheme. In addition to wages, tax is withheld on a number of other transactions including payments for goods and services, utility bills, cash withdrawal from banks, and imports from other countries. The withheld tax can be adjusted against the tax liability at the time of filing of returns. The firms that withhold tax are required to file a statement with the FBR indicating the transactions and the tax withheld thereon. The scheme has some elements of third party reporting, though it does not provide complete information on the tax base itself as is the case with

[91] The cutoff choice reflects strongly skewed size distribution of firms. The 75th percentile corresponds to an annual turnover of Rs. 6.6 million (US $ 65,600). Compared to this the 50th percentile firm has a turnover of US$ 1,860 only.

tax withholding on wages. The dummy variable takes the value 1 if withheld tax of a firm as a percentage of its taxable income in pre-reform years is more than 75th percentile.

Table A.2: Robustness of Bunching Estimates

Panel A: Varying the Order of Polynomial

Order of Polynomial	3	4	5	6	7
High-rate Firms, 2006/07/09	4.28	3.92	4.44	6.05	5.53
	(.1)	(.1)	(.1)	(.1)	(.1)
Low-rate Firms, 2006/07/09	1.93	2.04	2	2.47	2.5
	(.2)	(.2)	(.2)	(.2)	(.2)
High-rate Firms, 2010	2.55	2.25	2.05	1.48	1.41
	(.2)	(.2)	(.2)	(.2)	(.1)

Panel B: Varying the Number of Excluded Bins

Number of Excluded Bins	1	2	3	4	5
High-rate Firms, 2006/07/09	1.83	2.57	3.65	4.44	2.23
	(.1)	(.1)	(.1)	(.1)	(.1)
Low-rate Firms, 2006/07/09	1.7	2	2.01	1.48	1.45
	(.1)	(.2)	(.2)	(.3)	(.3)
High-rate Firms, 2010	1.82	2.05	2.6	2.43	2.34
	(.1)	(.2)	(.2)	(.2)	(.3)

Notes: The table presents estimates of the excess mass b, for different specifications of the estimating equation (**??**), for the subsamples considered in table **??**. 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 **??** (4 bins on either side of the kink for high-rate firms in 2006/07/09, 2 bins otherwise). Panel B presents estimates for different choices of the excluded range ($1 - 5$ bins on either side of the kink), for the order of polynomial chosen as in table **??** ($q = 5$). Bootstrapped standard errors are shown in parantheses.

Table A.3: Estimating Output Evasion Responses

	Observed Responses		**Without Evasion**	**With Evasion**			
	Bunching (b)	**Profit Rate (Δp)**	**Output Elasticity (ε_y)**	**Evasion Rate Response**			
				$\varepsilon_y = 0$	$\varepsilon_y = 0.5$	$\varepsilon_y = 1$	$\varepsilon_y = 5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
High-rate Firms, 2006/07/09	4.44	1.0	133.3	67.6	67.4	67.1	67.5
	(0.1)	(0.03)	(3.8)	(2.0)	(2.0)	(2.0)	(2.0)
Low-rate Firms, 2006/07/09	2.00	0.4	34.3	17.6	17.3	17.1	15
	(0.2)	(0.04)	(3.0)	(1.7)	(1.7)	(1.7)	(1.7)
High-rate Firms, 2010	2.05	0.4	14.7	15.1	14.6	14.1	10.0
	(0.2)	(0.03)	(1.2)	(1.3)	(1.3)	(1.3)	(1.3)

Notes: This table presents bunching and elasticity estimates for the subsamples considered in panels A, B and D of figure **??**. Column (1) reproduces the bunching estimate b, based on estimating equation (**??**). 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 associated with b, based on the relationship $\Delta p = B/f_0\left(\tau_y/\tau_\pi\right) \simeq b \times binwidth$. 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. ε_y is estimated using the relationship $\Delta p = \left[c/y - c'\left(y\right)\right] dy/y \simeq \left(\tau_y^2/\tau_\pi\right)\varepsilon_y$. Columns (4)-(7) present estimates of the output evasion response as percentage of taxable profits (evasion *rate* responses), for the model with output evasion but no cost evasion, as presented in section **??**. This model allows for bunching to be driven by both output evasion and real output response. The evasion response estimates are based on $\Delta\hat{p} = \epsilon_y\left(y/\hat{y}\right)\left(\tau_y/\tau_\pi\right) - \left(1-\tau_y/\tau_\pi\right)\left(\tau_y/\tau_\pi\right)\left[d(y-\hat{y})/(\hat{y}-\hat{c})\right]$, assuming different real output elasticities $\varepsilon_y \in \{0, 0.5, 1, 5\}$). Bootstrapped standard errors are shown in parentheses.

Table A.4: Data Cleaning Steps

	Panel A: Sample Definition
Sample	**Definition**
Firms Reporting Profits & Turnover	Firms reporting profits Π, turnover y and incorporation date D. Based on Π and y, derive implied tax liablities $\tilde{T}^y$, $\tilde{T}^\Pi_H$ and $\tilde{T}^\Pi_L$ 4 (high and low profit rate).
Consistency Check I	Drop firm if reported and implied tax liability 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 missing, assign {H,L} based on y, D and capital K.
Consistency Check II	Drop firm if reported and implied taxpayer status 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 missing ; $\tilde{T}^y < \tilde{T}^\Pi$ and T^Π missing.

	Panel B: Sample Size		
Step	**Year**	**High-rate Firms**	**Low-rate Firms**
Raw data	2006/07/09	45,284	
	2008	21,445	
	2010	21,584	
Firms Reporting Profits & Turnover	2006/07/09	10,228	2,899
	2008	4,515	1,546
	2010	4,862	1,867
After Consistency Check I	2006/07/09	10,265	2,198
	2008	4,706	1,126
	2010	5,212	1,418
After Consistency Check II	2006/07/09	9,472	1,966
	2008	4,706	1,115
	2010	4,678	1,239

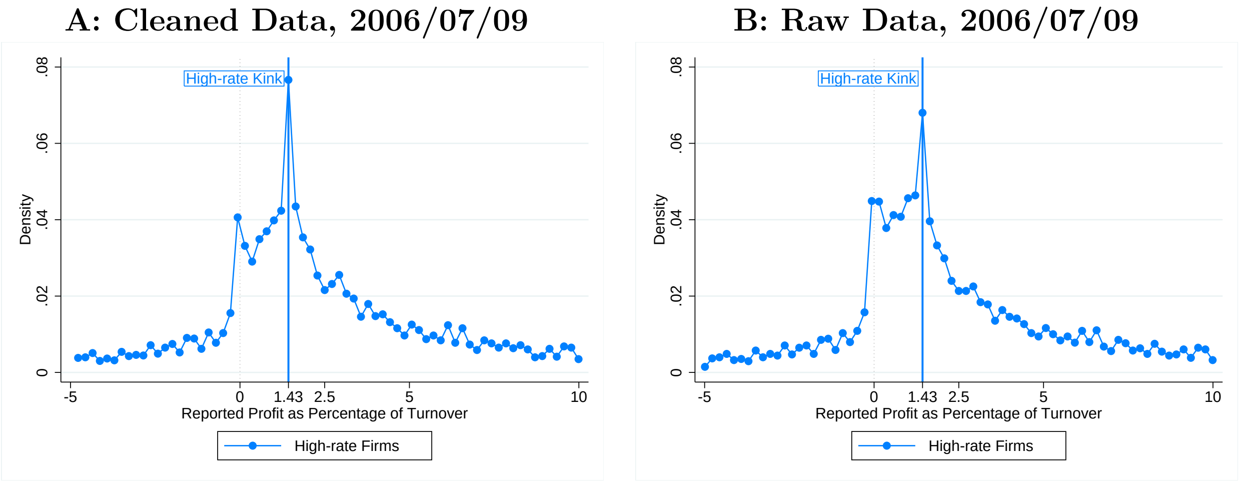
Notes: Panel A of this table explains the consistency checks applied to the data. For all consistency checks, a tolerance threshold of 5% is used. Panel B displays the sample size for different steps in the cleaning process. Capital K is equity plus retained earnings. Note that the implied turnover tax liability used for consistency check I is gross implied turnover tax liability minus net deductions (which are deduced from the tax liablity 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-net reductions)/turnover, for firms that report positive net reductions.

Table A.5: Comparison of Missing and Non-missing Observations

Panel A: Firms Reporting Profits and Turnover				
	N (1)	Median (2)	Mean (3)	SD (4)
Profits	17,358	0.1	-35.3	1624
Turnover	17,358	25.1	711.7	5579.6
Salary	7,865	8.8	63	265.6
Interest	9,726	0.9	84.8	901.2
Share of Low-rate Firms	17358		0.18	
Panel B: Firms Reporting Profits Only				
Profits	13,155	0	-69.8	2483.8
Salary	3,500	15	105.5	514
Interest	5,176	0.7	157.9	1472.3
Share of Low-rate Firms	11814		0.14	
Panel C: Firms Reporting Turnover Only				
Turnover	8,551	9.2	454.5	7399.1
Salary	3,078	5	37.8	271.8
Interest	3,767	0.7	40.2	259.9
Share of Low-rate Firms	8,551		0.27	

Notes: The table compares different samples of firms depending on wether or not they report profits and turnover. 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 million Pakistani Rupees (PKR).

Figure A.6: Bunching among High-rate Firms

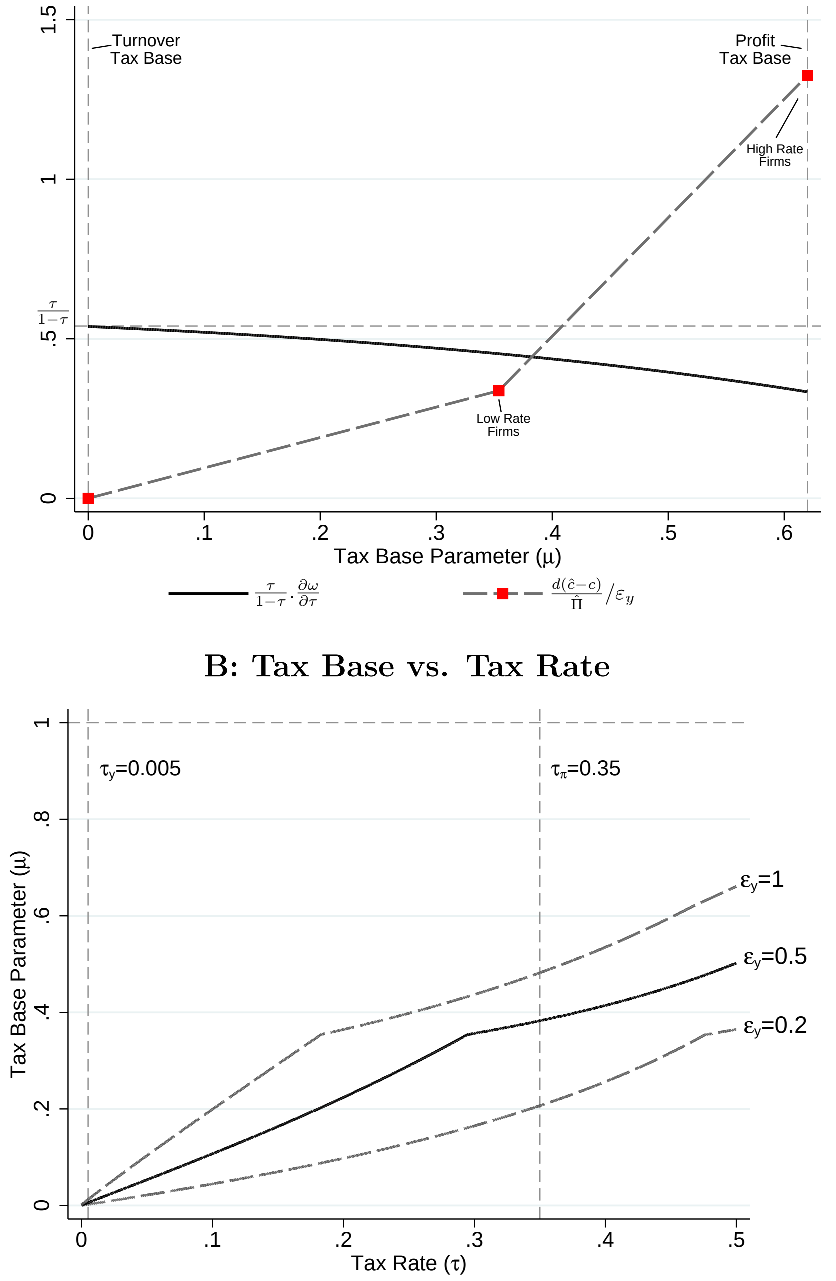

A: Cleaned Data, 2006/07/09

B: Raw Data, 2006/07/09

Notes: This figure shows bunching in the sample of high-rate firms in 2006/07/09, in the cleaned data (panel A, same as panel A in figure **??**), and the raw data (panel B). The raw data contains all observations that report turnover and profits, before consistency checks I and II are applied, and before firms reporting $\pi = 0$ are dropped. The graphs show the empirical density distribution of the profit rate (reported profit as percentage of turnover). The tax liablity for a firm with output y and cost $c(y)$ is $T\left[y, c\left(y\right)\right] = \max\left\{\tau_{\pi}\left[y - c\left(y\right)\right], \tau_{y} y\right\}$, where τ_{π} is the profit tax rate and τ_y is the turnover tax rate. The ratio of these two tax rates marks the kink at which firms move from the profit tax scheme (for profit rates above the kink) to the turnover tax scheme (for profit rates below the kink). For high-rate firms in 2006/07/09, $\tau_{\pi} = 0.35$ and $\tau_y = 0.005$, placing the kink at a profit rate of 1.43%. The kink point is marked by a vertical solid line. The zero profit point is marked by a vertical dotted line. The bin size is 0.214, chosen so that the kink point is a bin centre.

Figure A.7: Numerical Analysis of Optimal Tax Policy

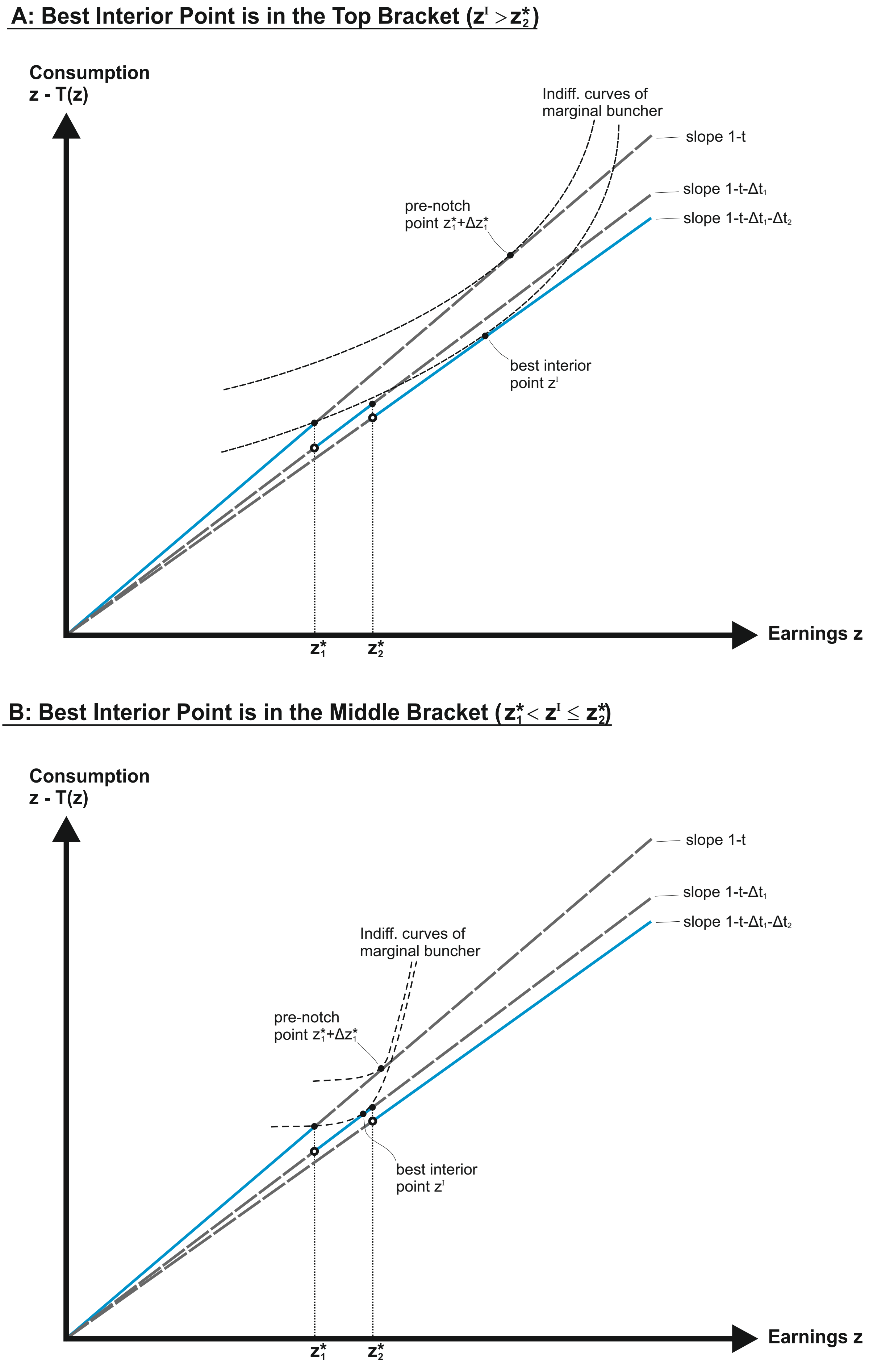
A: Optimal Tax Rule



B: Tax Base vs. Tax Rate

Notes: This figure considers the optimal tax policy implied by the optimal tax rule (**??**) using a distortionary profit tax as a benchmark. We rescale our tax base parameter μ so that the implied tax wedge ω equals 17% – the mean effective tax rate on profits reported in Gruber & Rauh (2007) – implying $\mu = 0.62$ when $\tau_\pi = 0.35$. The solid black curve in panel A plots the left-hand side of the optimal tax rule equation as a function of the rescaled μ for $\varepsilon_y = 0.5$ and $\tau_\pi = 0.35$. The three red markers on the dashed gray curve show respectively the right-hand side of the optimal tax rule at $\mu = 0$, at $\mu = (0.2/0.35) \times 0.62$ based on the evasion rate response estimated for the low-rate firms, and at $\mu = 0.62$ based on the evasion rate response estimated for the high-rate firms. By extrapolating between these three estimates, we find that the optimal tax base implied by the tax rule equals $\mu = 0.383$ as compared to $\mu = 0.578$ for a completely nondistortionary profit tax. In panel B we replicate the exercise to find the optimal tax base as a function of the tax rate for three different levels of the output elasticity.

Figure A.8
Multiple-Notch Setting Where Bunchers Jump Two Notches



A: Best Interior Point is in the Top Bracket ($z^I > z_2^*$)

B: Best Interior Point is in the Middle Bracket ($z_1^* < z^I \leq z_2^*$)

FIGURE A.9

Triangular Missing Mass and the Shape of the Post-Notch Distribution

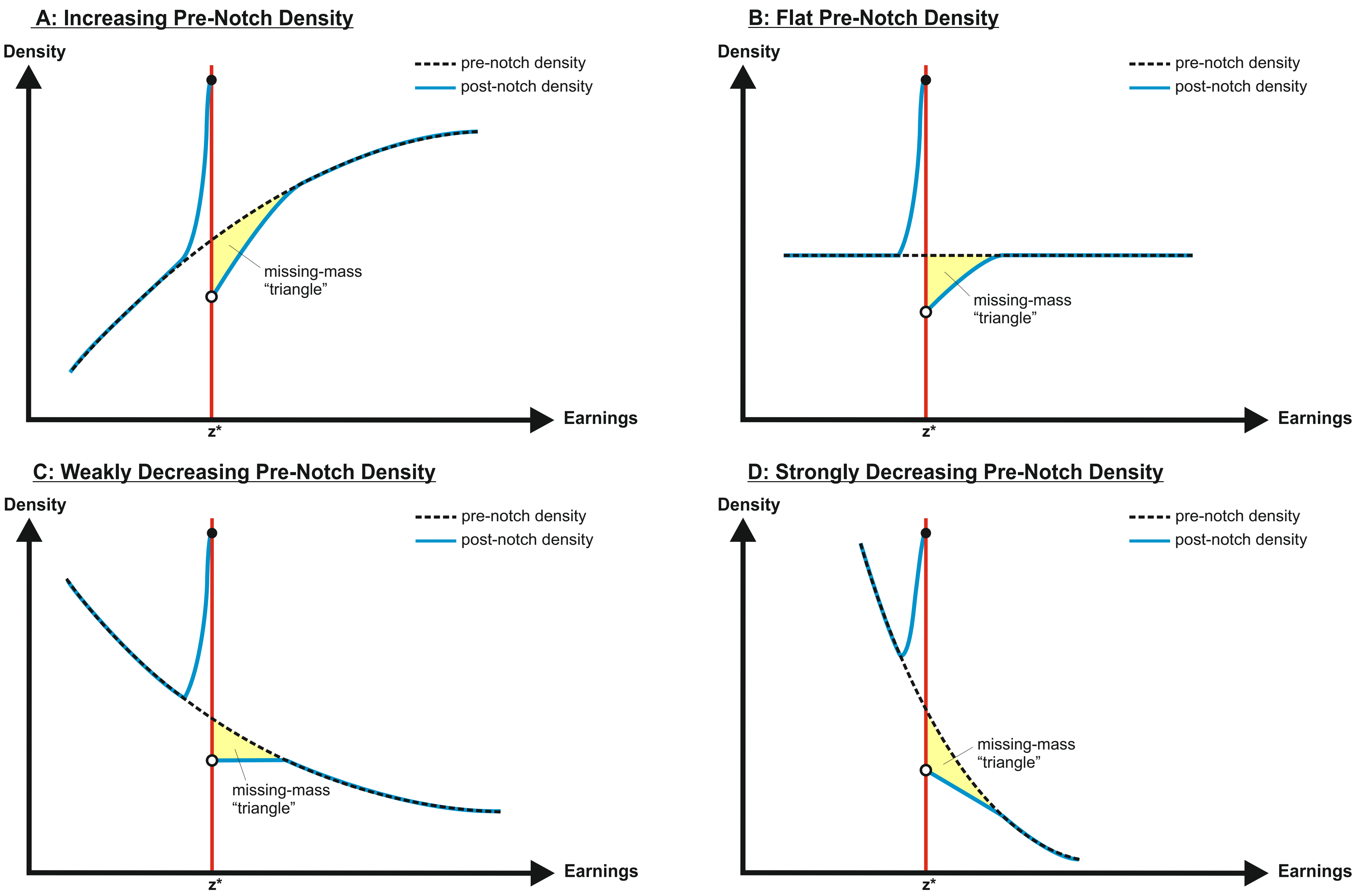
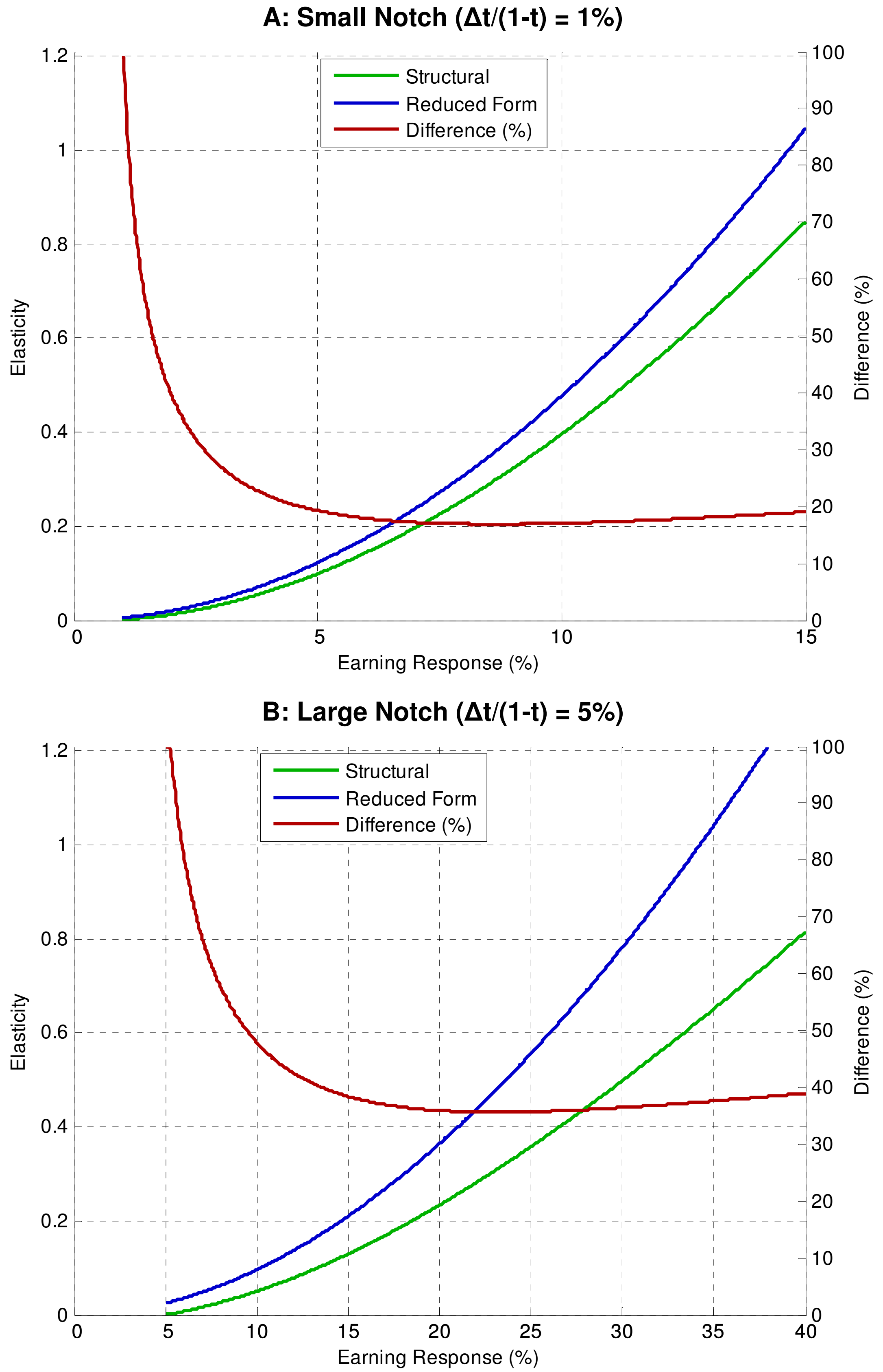


Figure A.10

Reduced-Form Approximation vs. True Structural Elasticity



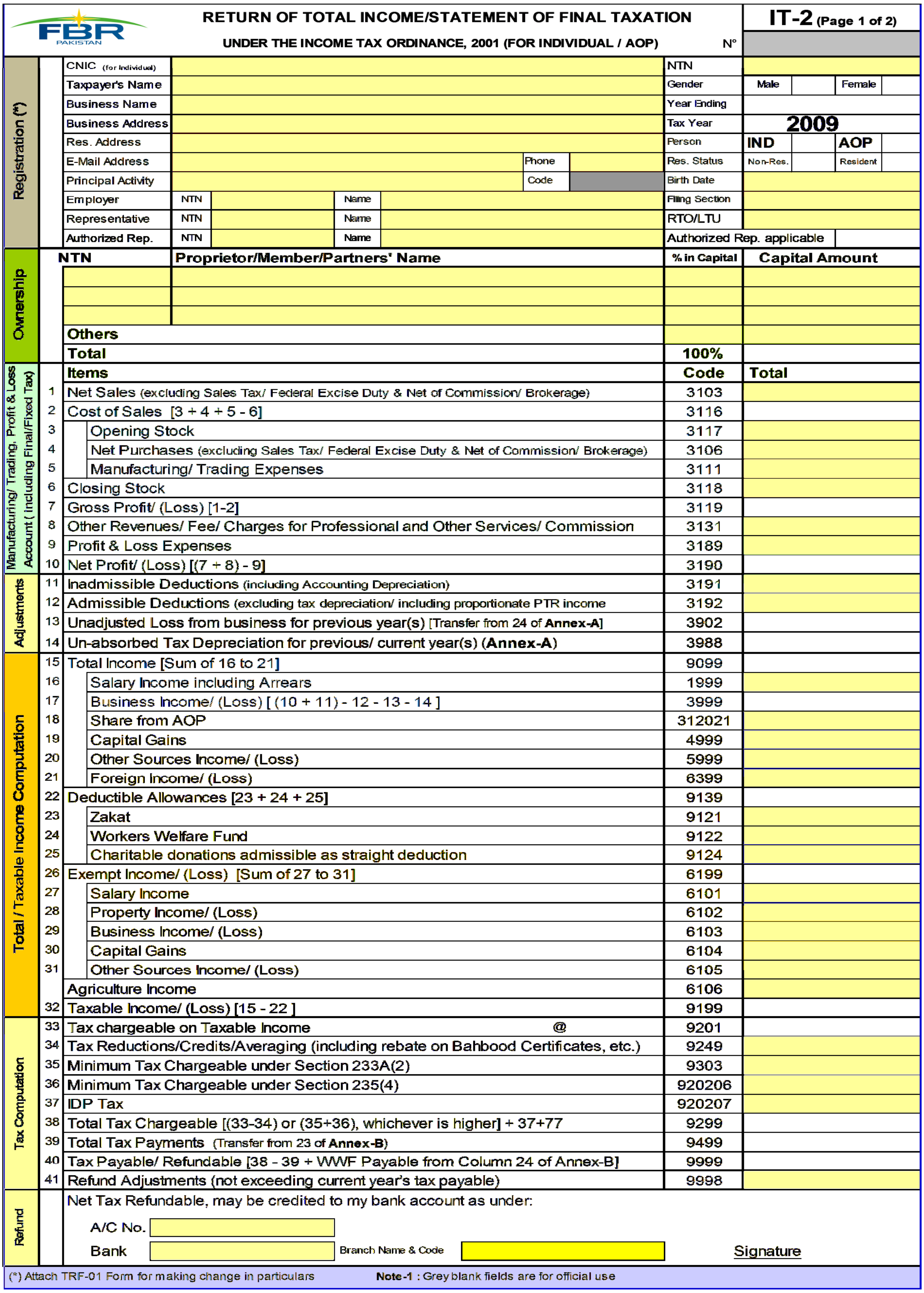
A: Small Notch (Δt/(1-t) = 1%)

B: Large Notch (Δt/(1-t) = 5%)

Notes: Assuming that true preferences are quasi-linear, the green graph shows the true structural elasticity from eq. (5), the blue graph shows the reduced-form approximation from eq. (12), and the red graph shows the percentage difference between the two as (reduced-form – structural)/reduced-form. Elasticity levels are depicted on the left y-axis while the elasticity difference is depicted on the right y-axis. Absolute bias is increasing in the earnings response and in the size of the notch. Relative bias is overall decreasing in the earnings response and increasing in the size of the notch.

Figure A.11

Personal Income Tax Return in Pakistan (Tax Year July 2008 – June 2009)



FBR PAKISTAN

RETURN OF TOTAL INCOME/STATEMENT OF FINAL TAXATION

UNDER THE INCOME TAX ORDINANCE, 2001 (FOR INDIVIDUAL / AOP) N°

IT-2 (Page 1 of 2)

Registration (*)	CNIC (for Individual)					NTN			
	Taxpayer's Name					Gender	Male		Female
	Business Name					Year Ending			
	Business Address					Tax Year	2009		
	Res. Address					Person	IND		AOP
	E-Mail Address			Phone		Res. Status	Non-Res.		Resident
	Principal Activity			Code		Birth Date			
	Employer	NTN		Name		Filing Section			
	Representative	NTN		Name		RTO/LTU			
	Authorized Rep.	NTN		Name		Authorized Rep. applicable			

Ownership	**NTN**	**Proprietor/Member/Partners' Name**	**% in Capital**	**Capital Amount**
	Others			
	Total		**100%**	

Manufacturing/ Trading, Profit & Loss Account (including Final/Fixed Tax)		**Items**	**Code**	**Total**
	1	Net Sales (excluding Sales Tax/ Federal Excise Duty & Net of Commission/ Brokerage)	3103	
	2	Cost of Sales [3 + 4 + 5 - 6]	3116	
	3	Opening Stock	3117	
	4	Net Purchases (excluding Sales Tax/ Federal Excise Duty & Net of Commission/ Brokerage)	3106	
	5	Manufacturing/ Trading Expenses	3111	
	6	Closing Stock	3118	
	7	Gross Profit/ (Loss) [1-2]	3119	
	8	Other Revenues/ Fee/ Charges for Professional and Other Services/ Commission	3131	
	9	Profit & Loss Expenses	3189	
	10	Net Profit/ (Loss) [(7 + 8) - 9]	3190	
Adjustments	11	Inadmissible Deductions (including Accounting Depreciation)	3191	
	12	Admissible Deductions (excluding tax depreciation/ including proportionate PTR income	3192	
	13	Unadjusted Loss from business for previous year(s) [Transfer from 24 of **Annex-A**]	3902	
	14	Un-absorbed Tax Depreciation for previous/ current year(s) **(Annex-A)**	3988	
Total / Taxable Income Computation	15	Total Income [Sum of 16 to 21]	9099	
	16	Salary Income including Arrears	1999	
	17	Business Income/ (Loss) [(10 + 11) - 12 - 13 - 14]	3999	
	18	Share from AOP	312021	
	19	Capital Gains	4999	
	20	Other Sources Income/ (Loss)	5999	
	21	Foreign Income/ (Loss)	6399	
	22	Deductible Allowances [23 + 24 + 25]	9139	
	23	Zakat	9121	
	24	Workers Welfare Fund	9122	
	25	Charitable donations admissible as straight deduction	9124	
	26	Exempt Income/ (Loss) [Sum of 27 to 31]	6199	
	27	Salary Income	6101	
	28	Property Income/ (Loss)	6102	
	29	Business Income/ (Loss)	6103	
	30	Capital Gains	6104	
	31	Other Sources Income/ (Loss)	6105	
		Agriculture Income	6106	
	32	Taxable Income/ (Loss) [15 - 22]	9199	
Tax Computation	33	Tax chargeable on Taxable Income @	9201	
	34	Tax Reductions/Credits/Averaging (including rebate on Bahbood Certificates, etc.)	9249	
	35	Minimum Tax Chargeable under Section 233A(2)	9303	
	36	Minimum Tax Chargeable under Section 235(4)	920206	
	37	IDP Tax	920207	
	38	Total Tax Chargeable [(33-34) or (35+36), whichever is higher] + 37+77	9299	
	39	Total Tax Payments (Transfer from 23 of **Annex-B**)	9499	
	40	Tax Payable/ Refundable [38 - 39 + WWF Payable from Column 24 of Annex-B]	9999	
	41	Refund Adjustments (not exceeding current year's tax payable)	9998	

Refund

Net Tax Refundable, may be credited to my bank account as under:

A/C No.

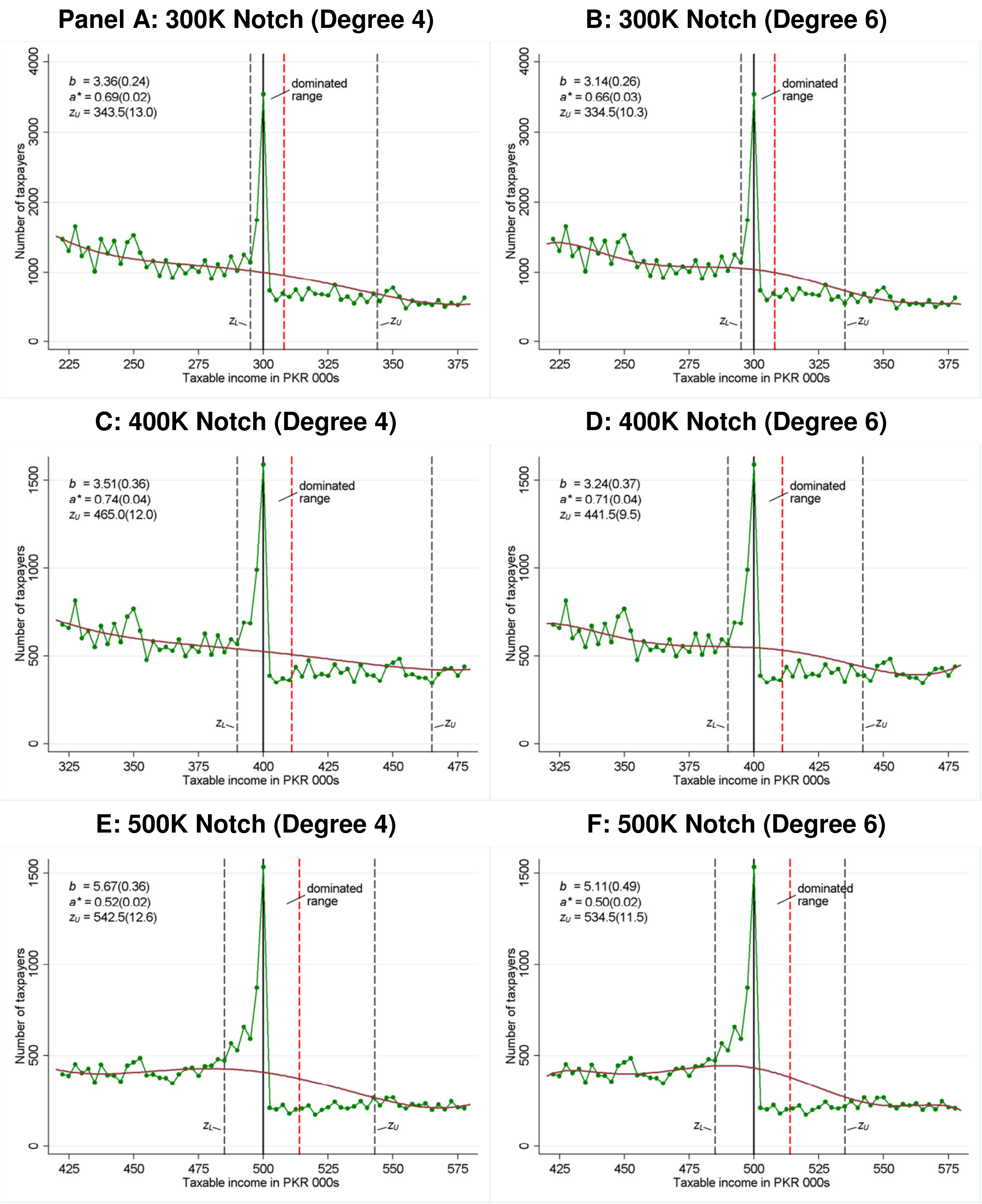
Bank Branch Name & Code Signature

(*) Attach TRF-01 Form for making change in particulars **Note-1** : Grey blank fields are for official use

Notes: the taxable income measure that applies to the tax rate schedule in Figure V is reported in cell 32, tax liability (= taxable income x relevant average tax rate in Figure V) is reported in cell 33, and final tax payable (tax liability with various adjustments and net of income tax withholding) is reported in cell 40. The filed return is subject to a validation check that uncovers internal inconsistencies (e.g., between taxable income in cell 32 and tax liability in cell 33). We observe such inconsistencies, which provide indicators of misperception of either the tax rate schedule or the tax return itself (as the tax computation cells 33-41 create scope for confusion, especially for those subject to withholding).

Figure A.12

Empirical and Counterfactual Distributions around Notches:
Self-Employed Individuals (Non-Rounder Sample)



Panel A: 300K Notch (Degree 4)

B: 300K Notch (Degree 6)

C: 400K Notch (Degree 4)

D: 400K Notch (Degree 6)

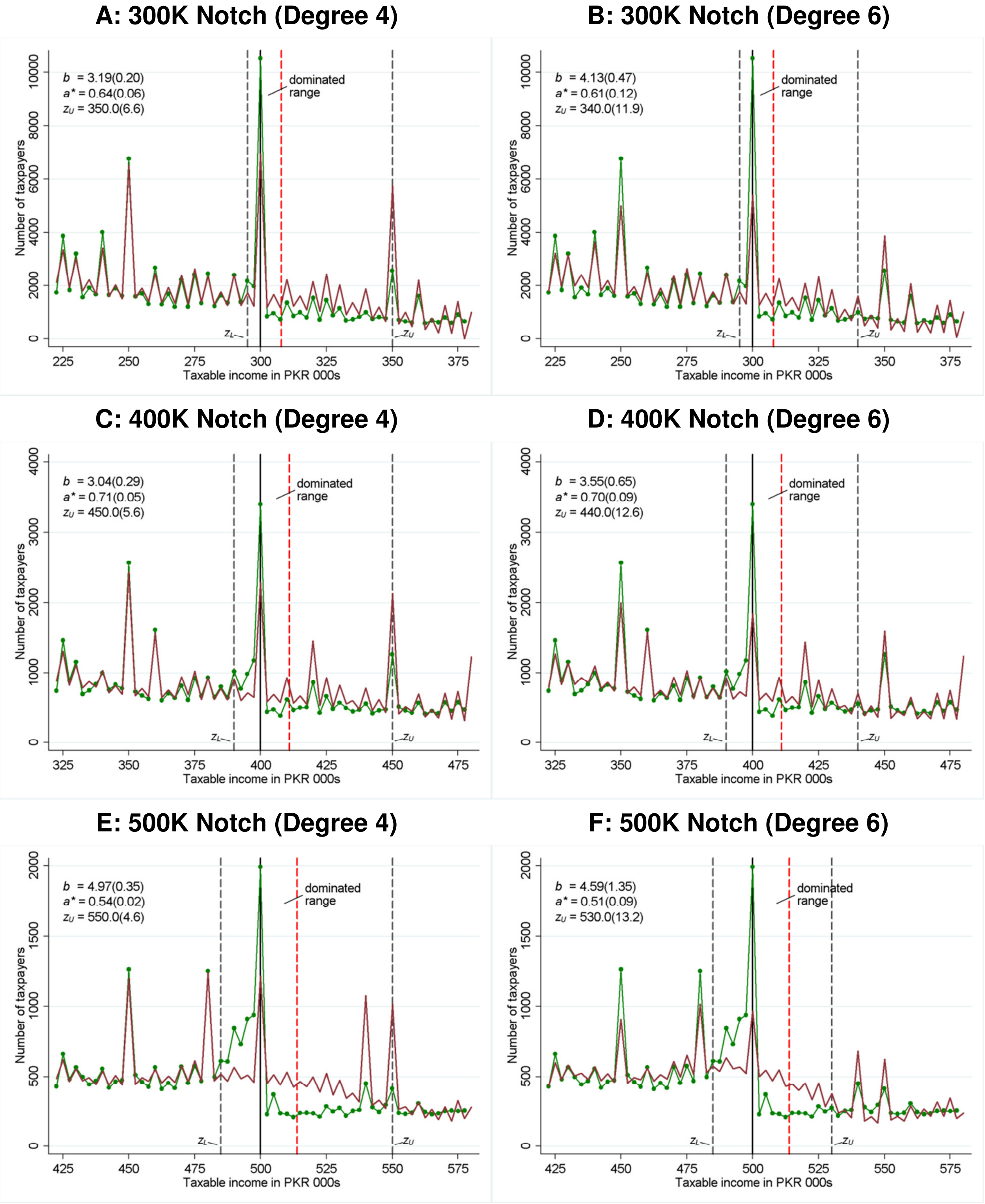
E: 500K Notch (Degree 4)

F: 500K Notch (Degree 6)

Notes: the figure shows the empirical distribution of taxable income (dotted green graph) and the counterfactual distribution (solid brown graph) for self-employed individuals (non-rounder sample) from 2006-09. The counterfactual is estimated for each notch separately by fitting either a fourth-order polynomial (left panels) or a sixth-order polynomial (right panels) to the empirical distribution, excluding data around the notch, as specified in equation (13). Notch points are marked by solid black lines, upper bounds of dominated regions are marked by dashed red lines, and excluded ranges [z_L,z_U] are marked by dashed black lines. Bunching b is excess mass in the excluded range below the notch (in proportion to the average counterfactual frequency in the dominated range), a^* is the share of individuals in the dominated range who are unresponsive, and the upper bound of the excluded range z_U has been estimated to ensure that missing mass equals bunching mass. Standard errors are shown in parentheses. The estimates of b and a^* are very robust to polynomial degree, while the estimates of z_U are slightly more sensitive.

Figure A.13

Empirical and Counterfactual Distributions around Notches:
Self-Employed Individuals (Full Sample)



A: 300K Notch (Degree 4)

B: 300K Notch (Degree 6)

C: 400K Notch (Degree 4)

D: 400K Notch (Degree 6)

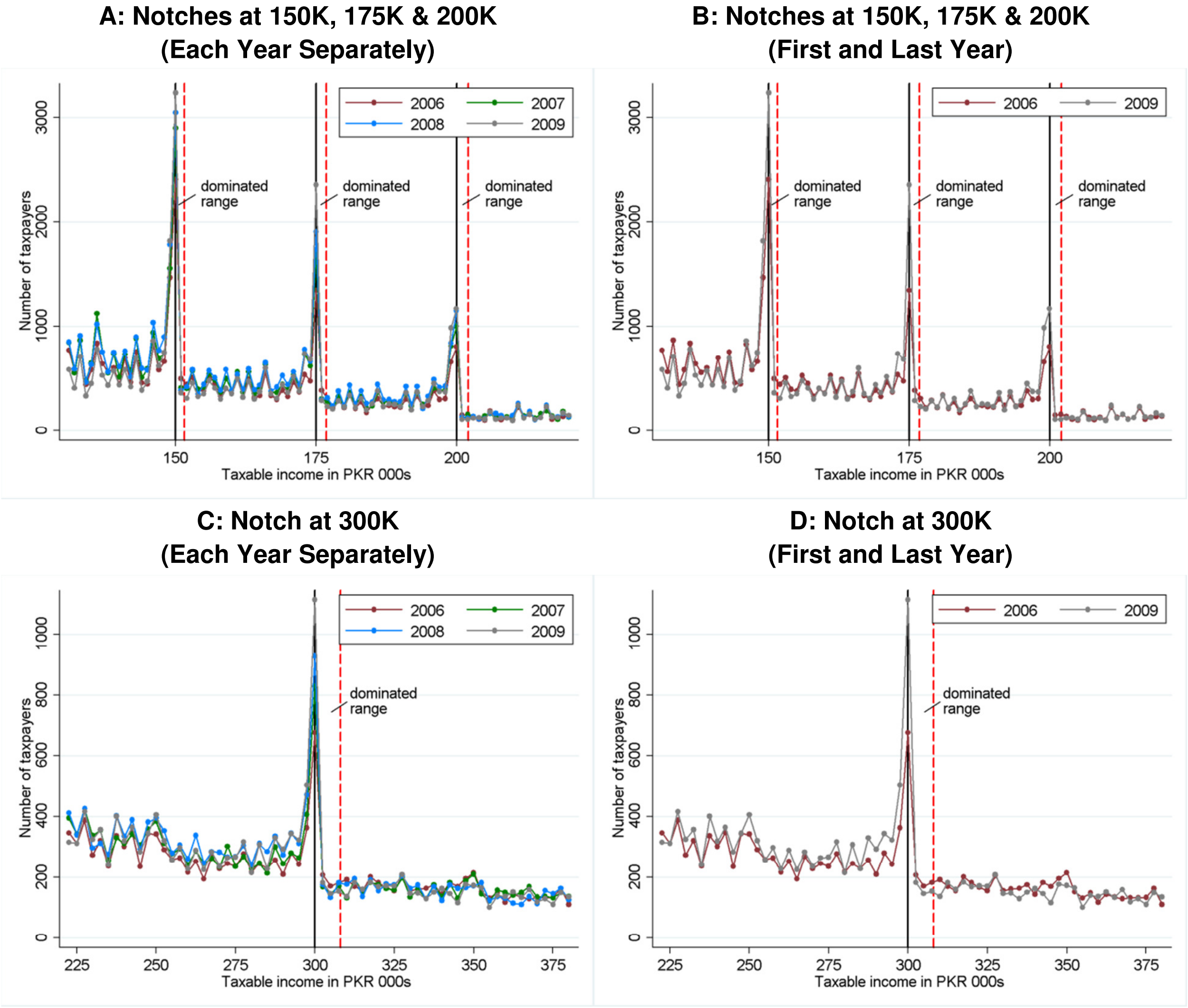
E: 500K Notch (Degree 4)

F: 500K Notch (Degree 6)

Notes: the figure shows the empirical distribution of taxable income (dotted green graph) and the counterfactual distribution (solid brown graph) for self-employed individuals (full sample) from 2006-09. The counterfactual is estimated for each notch separately by fitting either a fourth-order polynomial (left panels) or a sixth-order polynomial (right panels) with round-number fixed effects to the empirical distribution, excluding data around the notch, as specified in equation (14). Notch points are marked by solid black lines, upper bounds of dominated regions are marked by dashed red lines, and excluded ranges [z_L,z_U] are marked by dashed black lines. Bunching *b* is excess mass in the excluded range below the notch (in proportion to the average counterfactual frequency in the dominated range), *a** is the share of individuals in the dominated range who are unresponsive, and the upper bound of the excluded range z_U has been estimated to ensure that missing mass equals bunching mass. Standard errors are shown in parentheses. Estimates for the full sample tend to be more sensitive than for the non-rounder sample (but not frictions *a**, which is always very robust).

Figure A.14

Dynamics of Bunching and Dominated Behaviour:
Self-Employed Individuals (Non-Rounder Sample)



A: Notches at 150K, 175K & 200K (Each Year Separately)

B: Notches at 150K, 175K & 200K (First and Last Year)

C: Notch at 300K (Each Year Separately)

D: Notch at 300K (First and Last Year)

Notes: the figure shows the empirical distribution of taxable income in 2006 (dotted red graph), 2007 (dotted green graph), 2008 (dotted blue graph), and 2009 (dotted grey graph) for self-employed individuals (non-rounder sample). Left panels show all four years together, while right panels show only the first and last years to make the graphs clearer. Notch points are marked by solid black lines and upper bounds of dominated regions are marked by dashed red lines. The graphs show that bunching is increasing while dominated behavior is falling over time.

Bibliography

Allingham, Michael G., & Sandmo, Agnar. 1972. Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics*, **1**, 323–338.

Andreoni, James, Erard, Brian, & Feinstein, Jonathan. 1998. Tax complicance. *Journal of Economic Literature*, **36**, 818–860.

Auerbach, Alan J. 2002. Taxation and Corporate Financial Policy. *Chap. 8, pages 1251–1292 of:* Auerbach, Alan J., & Feldstein, Martin (eds), *Handbook of Public Economics, Volume 3.*

Auerbach, Alan J., Devereux, Michael P., & Simpson, Helen. 2010. Taxing Corporate Income. *Chap. 9 of: Dimensions of Tax Design: the Mirrlees Review.* Oxford University Press.

Bach, Laurent. 2012 (January). *Tax Collection and Private Governance: Evidence from French Businesses.* Working Paper.

Bachas, Pierre, & Jensen, Anders. 2013. *Information Trails, Tax Compliance and Development.* Mimeo.

Baunsgaard, Thomas, & Keen, Michael. 2010. Tax revenue and (or?) trade liberalization. *Journal of Public Economics*, **94**(9-10), 563–577.

Besley, Timothy, & Persson, Torsten. 2011. *Pillars of Prosperity: The Political Economy of Development Clusters.* Princeton University Press.

Besley, Timothy, & Persson, Torsten. 2013. *Taxation and Development.* Chapter for Handbook of Public Economics, Vol. 5. forthcoming.

Blundell, Richard, & Hoynes, Hilary W. 2004. Has' In-Work'Benefit Reform Helped the Labor Market? *Pages 411–460 of: Seeking a Premier Economy: The Economic Effects of British Economic Reforms, 1980-2000.* University of Chicago Press.

Blundell, Richard, & Shephard, Andrew. 2012. Employment, hours of work and the optimal taxation of low-income families. *The Review of Economic Studies*, **79**(2), 481–510.

Boadway, Robin, & Sato, Motohiro. 2009. Optimal Tax Design and Enforcement with an Informal Sector. *American Economic Journal: Economic Policy*, **1**(1), 1–27.

Cage, Julia, & Gadenne, Lucie. 2012. *The Fiscal Cost of Trade Liberalization.* Working paper.

Chetty, Raj. 2009. Is the Taxable Income Elasticity Sufficient to Calculate Deadweight Loss? The Implications of Evasion and Avoidance. *American Economic Journal: Economic Policy*, **1**(2), 31–52.

Chetty, Raj. 2012. Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply. *Econometrica*, **80**(3), 969–1018.

Chetty, Raj, & Saez, Emmanuel. 2013. Teaching the tax code: Earnings responses to an experiment with EITC recipients. *American Economic Journal: Applied Economics*, **5**(1), 1–31.

Chetty, Raj, Looney, Adam, & Kroft, Kory. 2009. Salience and Taxation: Theory and Evidence. *American Economic Review*, **99**(4), 1145–77.

Chetty, Raj, Friedman, John, Olsen, Tore, & Pistaferri, Luigi. 2011. Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records. *Quarterly Journal of Economics*, **126**, 749–804.

Christensen, Kevin, Cline, Robert, & Neubig, Tom. 2001. Total Corporate Taxation:" Hidden," Above-the-Line, Non-Income Taxes. *National Tax Journal*, **54**(3), 495–506.

de Paula, Aureo, & Scheinkman, Jose A. 2010. Value-Added Taxes, Chain Effects, and Informality. *American Economic Journal: Macroeconomics*, **2**(4), 195–221.

Devereux, Michael, Liu, Li, & Loretz, Simon. 2013. *The Elasticity of Corporate Taxable Income: New Evidence from UK Tax Records.* Oxford University Centre for Business Taxation Working Paper 12/23.

Devereux, Michael P., & Sorensen, Peter Birch. 2006. *The Corporate Income Tax: International Trends and Options for Fundamental Reform.* Economic Papers. European Commission.

Dharmapala, Dhammika, Slemrod, Joel, & Wilson, John Douglas. 2011. Tax Policy and the Missing Middle: Optimal Tax Remittance with Firm-Level Administrative Costs. *Journal of Public Economics*, **72**, 1036–1047.

Diamond, Peter A, & Mirrlees, James A. 1971. Optimal Taxation and Public Production I: Production Efficiency. *American Economic Review*, **61**(1), 8–27.

Dwenger, Nadja, & Steiner, Viktor. Profit Taxation and the Elasticity of the Corporate Income Tax Base: Evidence from German Corporate Tax Return Data. *National Tax Journal*, **65**(March), 117–150.

Eissa, Nada, & Hoynes, Hillary. 2000. Tax and transfer policy, and family formation: Marriage and cohabitation. *University of California Davis. Unpublished.*

Emran, M. Shahe, & Stiglitz, Joseph E. 2005. On Selective Indirect Tax Reform in Developing Countries. *Journal of Public Economics*, **89**, 599–623.

Ernst & Young. 2013. *Worldwide Corporate Tax Guide.* Tech. rept. Ernst & Young.

FBR. 2010. *Salient Features For The Budget 2010-11.* Tech. rept. Federal Board of Revenue, Pakistan.

Feldstein, Martin. 1995. Behavioral Responses to Tax Rates: Evidence from the Tax Reform Act of 1986. *Ameican Economic Review*, **81**, 674–680.

Feldstein, Martin. 1999. Tax Avoidance And The Deadweight Loss Of The Income Tax. *The Review of Economics and Statistics*, **81**(4), 674–680.

Fuest, Clemens, & Riedel, Nadine. 2009. Tax evasion, tax avoidance and tax expenditures in developing countries: A review of the literature. *Report prepared for the UK Department for International Development (DFID).*

Goolsbee, Austan. 1998. Taxes, organizational form, and the deadweight loss of the corporate income tax. *Journal of Public Economics*, **69**(1), 143 – 152.

Goolsbee, Austan. 2004. The impact of the corporate income tax: evidence from state organizational form data. *Journal of Public Economics*, **88**, 2283 – 2299.

Gordon, Roger, & Li, Wei. 2009. Tax structures in developing countries: Many puzzles and a possible explanation. *Journal of Public Economics*, **93**(7-8), 855–866.

Gordon, Roger H., & MacKie-Mason, Jeffrey K. 1994. Tax Distortions to the Choice of Organizational Form. Jan.

Gordon, Roger H., & MacKie-Mason, Jeffrey K. 1997. How Much Do Taxes Discourage Incorporation? *Journal of Finance*, **52**(2), 477–505.

Gruber, Jon, & Saez, Emmanuel. 2002. The elasticity of taxable income: evidence and implications. *Journal of Public Economics*, **84**(1), 1 – 32.

Gruber, Jonathan, & Rauh, Joshua. 2007. How Elastic is the Corporate Income Tax Base? *In:* Hines, James R., Auerbach, Alan, & Slemrod, Joel (eds), *Taxing corporate income in the 21st century.* Cambridge University Press.

Gruber, Jonathan, & Wise, David A. 2008. *Social security and retirement around the world.* University of Chicago Press.

Hassett, Kevin A., & Hubbard, R. Glenn. 2002. Tax Policy and Business Investment. *Chap. 9, pages 1293–1343 of:* Auerbach, Alan J., & Feldstein, Martin (eds), *Handbook of Public Economics, Volume 3.*

Kawano, Laura, & Slemrod, Joel. 2012. *The Effect of Tax Rates and Tax Bases on Corporate Tax Revenues: Estimates With New Measures of the Corporate Tax Base.* NBER Working Paper 18440.

Keen, Michael. 2008. VAT, Tariffs, and Withholding: Border Taxes and Informality in Developing Countries. *Journal of Public Economics*, **92**, 1892–1906.

Keen, Michael. 2013. *Targeting, Cascading and Indirect Tax Design.* IMF Working Paper 13/57.

Kleven, Henrik J., & Waseem, Mazhar. 2013. Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan. *Quarterly Journal of Economics*, **128**, 669–723.

Kleven, Henrik Jacobsen, & Schultz, Esben Anton. 2011 (Aug.). *Estimating Taxable Income Responses using Danish Tax Reforms.* EPRU Working Paper Series 2011-02. Economic Policy Research Unit (EPRU), University of Copenhagen. Department of Economics.

Kleven, Henrik Jacobsen, Kreiner, Claus Thustrup, & Saez, Emmanuel. 2009. *Why Can Modern Governments Tax So Much? An Agency Model of Firms as Fiscal Intermediaries.* NBER Working paper 15218.

Kleven, Henrik Jacobsen, Knudsen, Martin B., Kreiner, Claus Thustrup, Pedersen, Sï¿œren, & Saez, Emmanuel. 2011. Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark. *Econometrica*, **79**(3), 651–692.

Kopczuk, Wojciech. 2005. Tax bases, tax rates and the elasticity of reported income. *Journal of Public Economics*, **89**(11-12), 2093–2119.

Kopczuk, Wojciech. 2012. *The Polish business "fl at" tax and its effect on reported incomes: a Pareto improving tax reform?* Working Paper, Columbia University.

Kopczuk, Wojciech, & Slemrod, Joel. 2006. Putting Firms into Optimal Tax Theory. *American Economic Review Papers and Proceedings*, **96**(2), 130–134.

Kumler, Todd, Verhoogen, Eric, & Frías, Judith. 2012. *Enlisting Workers in Monitoring Firms: Payroll Tax Compliance in Mexico.* Working Paper, Columbia University.

Manoli, Dayanand S, & Weber, Andrea. 2011. Nonparametric evidence on the effects of financial incentives on retirement decisions.

Pomeranz, Dina. 2013. *No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax.* Working Paper, Harvard Business School.

Ramnath, Shanthi. 2013. Taxpayers Responses' to Tax-Based Incentives for Retirement Savings: Evidence from the Savers' Credit Notch. *Journal of Public Economics.*

Saez, Emmanuel. 2004. Reported Incomes and Marginal Tax Rates, 1960-2000: Evidence and Policy Implications. *Pages 117–174 of: Tax Policy and the Economy, Volume 18.* NBER Chapters. National Bureau of Economic Research, Inc.

Saez, Emmanuel. 2010. Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy*, **2**(3), 180–212.

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

Sallee, James M, & Slemrod, Joel. 2012. Car notches: Strategic automaker responses to fuel economy policy. *Journal of Public Economics.*

Shaw, Jonathan, Slemrod, Joel, & Whiting, John. 2010. Administration and compliance. *Dimensions of Tax Design: the Mirrlees Review, J. Mirrlees, S. Adam, T. Besley, R. Blundell, S. Bond, R. Chote, M. Gammie, P. Johnson, G. Myles and J. Poterba (eds), Oxford University Press.*

Slemrod, Joel. 1998. Methodological Issues in Measuring and Interpreting Taxable Income Elasticities. *National Tax Journal*, **51**(4), 773–788.

Slemrod, Joel. 2007. Cheating Ourselves: The Economics of Tax Evasion. *Journal of Economic Perspectives*, **21**, 25–48.

Slemrod, Joel. 2010. *Buenas Notches: Lines and Notches in Tax System Design.*

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

Slemrod, Joel, & Weber, Caroline. 2012. Evidence of the invisible: toward a credibility revolution in the empirical analysis of tax evasion and the informal economy. *International Tax and Public Finance*, **19**(1), 25–53.

Slemrod, Joel, & Yitzhaki, Shlomo. 2002. Tax avoidance, evasion, and administration. *Chap. 22, pages 1423–1470 of:* Auerbach, A. J., & Feldstein, M. (eds), *Handbook of Public Economics.* Handbook of Public Economics, vol. 3. Elsevier.

Slemrod, Joel, Blumenthal, Marsha, & Christian, Charles. 2001. Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota. *Journal of Public Economics*, **79**(3), 455–483.

World Bank. 2009. *Pakistan Tax Policy Report:Tapping Tax Bases for Development.* Tech. rept. World Bank.

Yelowitz, Aaron S. 1995. The Medicaid notch, labor supply, and welfare participation: Evidence from eligibility expansions. *The Quarterly Journal of Economics*, **110**(4), 909–939.

Yitzhaki, Shlomo. 1974. Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics*, **3**, 201–202.