# USING NOTCHES TO UNCOVER OPTIMIZATION FRICTIONS AND STRUCTURAL ELASTICITIES: THEORY AND EVIDENCE FROM PAKISTAN*

Henrik J. Kleven and Mazhar Waseem

We develop a framework for nonparametrically identifying optimization frictions and structural elasticities using notches—discontinuities in the choice sets of agents—introduced by tax and transfer policies. Notches create excess bunching on the low-tax side and missing mass on the high-tax side of a cutoff, and they are often associated with a region of strictly dominated choice that would have zero mass in a frictionless world. By combining excess bunching (observed response attenuated by frictions) with missing mass in the dominated region (frictions), it is possible to uncover the structural elasticity that would govern behavior in the absence of frictions and arguably capture long-run behavior. We apply our framework to tax notches in Pakistan using rich administrative data. While observed bunching is large and sharp, optimization frictions are also very large as the majority of taxpayers in dominated ranges are unresponsive to tax incentives. The combination of large observed bunching and large frictions implies that the frictionless behavioral response to notches is extremely large, but the underlying structural elasticity driving this response is nevertheless modest. This highlights the inefficiency of notches: by creating extremely strong price distortions, they induce large behavioral responses even when structural elasticities are small. *JEL* Codes: H31, J22, O12.

## I. 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

* We are grateful for constructive comments and suggestions from editors Lawrence Katz and Robert Barro, five anonymous referees, Rashid Amjad, Tony Atkinson, Oriana Bandiera, Tim Besley, Michael Best, Richard Blundell, Raj Chetty, Estelle Dauchy, Roger Gordon, Damon Jones, Adnan Khan, Wojciech Kopczuk, Erzo Luttmer, Ben Olken, Torsten Persson, Thomas Piketty, Dina Pomeranz, Imran Rasul, Emmanuel Saez, John Karl Schultz, Monica Singhal, Joel Slemrod, Johannes Spinnewijn, Jakob Egholt Søgaard, and numerous seminar participants. Financial support from the International Growth Centre (IGC), Pakistan Programme is gratefully acknowledged.

*The Quarterly Journal of Economics* (2013), 669–723. doi:10.1093/qje/qjt004.
Advance Access publication on February 1, 2013.

highly parametric setting or in a way that addresses the frictions only qualitatively. This article develops a framework for nonparametrically 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 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).[1]

To understand the key idea of the article, 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 earnings distribution on the low-tax side of the notch point.[2] 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

[1] Existing empirical studies have considered behavioral responses to notches in these various contexts, including the U.S. Medicaid notch (Yelowitz 1995), social security notches (Gruber and Wise 1999; Manoli and Weber 2011), the U.S. Saver's Credit notch (Ramnath 2009), the U.K. in-work benefit notch (Blundell and Hoynes 2004; Blundell and Shephard 2012), and car taxation notches (Sallee and Slemrod 2012).

[2] 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.

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 and 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.[3] Moreover, the issue of optimization frictions that is central to this article is likely to be at least as important in underdeveloped 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. Although the use of large administrative data sets 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 provide for the first time compelling evidence of behavioral responses to taxes for a developing economy.

[3] A recent survey of the literature on taxation and development is provided by Besley and Persson (2012).

Our main findings are the following. First, there is large and sharp excess bunching below 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% of wage earners and 50%–80% 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.

The article is organized as follows. Section II develops the theoretical framework and empirical methodology, Section III presents the Pakistan application, and Section IV concludes.

## II. Theory and Empirical Methodology

### II.A. 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

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

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

$$z = n(1-t)^e, \tag{2}$$

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 (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 (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 (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 later 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 $\Delta T$ is a pure notch, $\Delta t$ is a proportional tax notch, and 1(.) is an indicator for being above the cutoff.

Figure I 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 $z^I$.[4]

The basic idea in the empirical approach is that the width of the bunching segment $\Delta z^*$ (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 $\Delta z^*$, 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

[4] While the utility specification (1) eliminates income effects, the implication of such effects can be seen from Figure I. The total response to the notch $\Delta z^*$ 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 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.

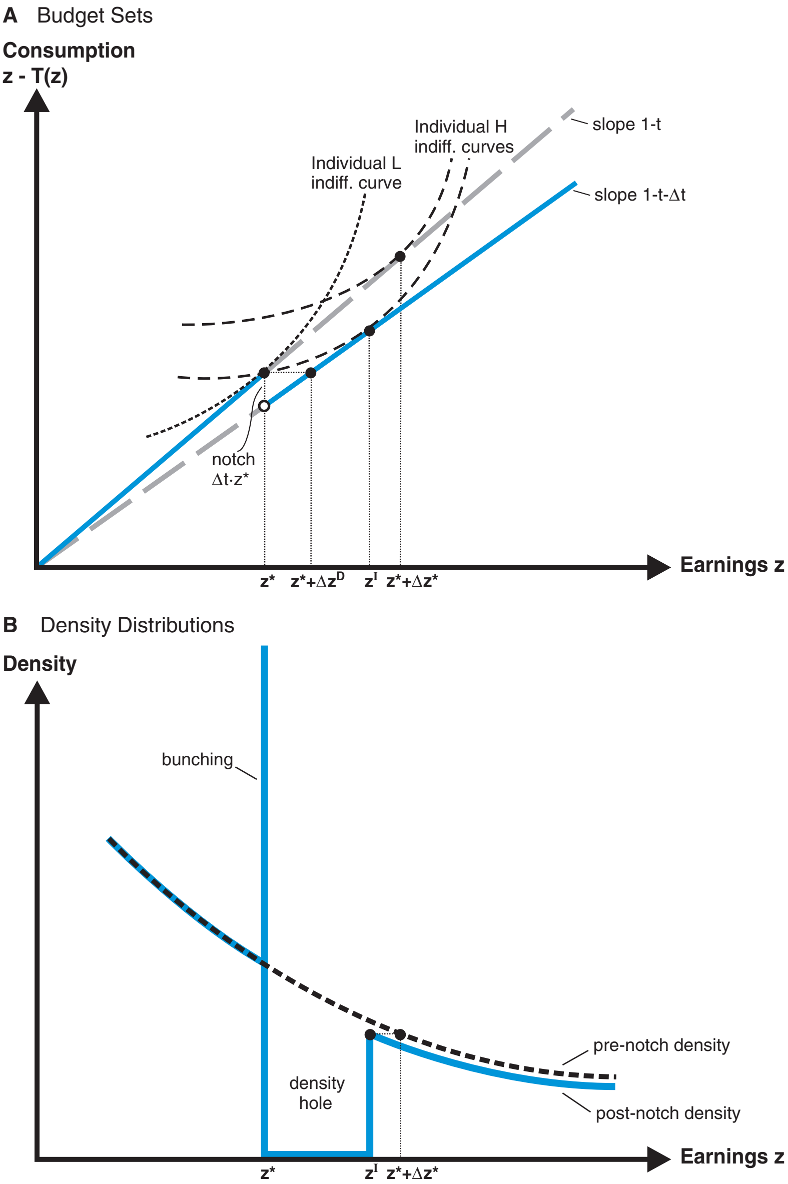


FIGURE I

Behavioral Responses to a Tax Notch

point $z^*$ and the best interior point $z^I$. At the notch point $z^*$, the utility level is given by

$$u^N = (1-t)z^* - \frac{n^* + \Delta n^*}{1+1/e}\left(\frac{z^*}{n^* + \Delta n^*}\right)^{1+1/e}. \tag{3}$$

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

$$u^I = \left(\frac{1}{1+e}\right)(n^* + \Delta n^*)(1 - t - \Delta t)^{1+e} - \Delta T. \tag{4}$$

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

$$\frac{1}{1+\Delta z^*/z^*}\left[1 + \frac{\Delta T/z^*}{1-t}\right] - \frac{1}{1+1/e}\left[\frac{1}{1+\Delta z^*/z^*}\right]^{1+1/e} - \frac{1}{1+e}\left[1 - \frac{\Delta t}{1-t}\right]^{1+e} = 0. \tag{5}$$

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 $\Delta z^*$ using bunching, it is useful to view the relationship (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 $\Delta z^*$ and observed values of the other arguments.[5]

There are two important points to note about the elasticity formula (5). First, as the compensated elasticity $e$ converges to zero (Leontief preferences), equation (5) implies

$$\lim_{e \to 0} \Delta z^* = \frac{\Delta T + \Delta t \cdot z^*}{1 - t - \Delta t} \equiv \Delta z^D. \tag{6}$$

Hence, under Leontief preferences, the bunching interval $\Delta z^*$ converges to the strictly dominated range $\Delta z^D$ in which taxpayers can increase both consumption and leisure by lowering earnings

[5] Formula (5) applies only to the case of *downward* budget set notches (increase in tax liability, decrease in transfers, etc.). The analysis of *upward* budget set notches is presented in the Online Appendix.

to the notch point.[6] The dominated range therefore represents a 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 (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: (1) 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. (2) 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 before, 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.1 of the Online Appendix. If the best interior point is located in the top bracket ($z^I > z_2^*$), the elasticity formula is equivalent to equation (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 (5) for $\Delta t \equiv \Delta t_1$ and $\Delta T \equiv \Delta T_1$.[7] Section II.C describes how we deal empirically with the possibility of bunchers jumping multiple notches.

The determination of the elasticity $e$ from equation (5) requires an estimate of the earnings response $\Delta z^*$. The model provides a relationship between the earnings response and estimable entities. Denoting excess bunching at the notch by $B$, we have

$$B = \int_{z^*}^{z^*+\Delta z^*} h_0(z)dz \approx h_0(z^*)\Delta z^*, \tag{7}$$

[6] The width of the dominated range $\Delta z^D$ is defined such that the earnings level $z^* + \Delta z^D$ ensures the same consumption as the notch point $z^*$, that is, $(1 - t - \Delta t)\left(z^* + \Delta z^D\right) - \Delta T = (1 - t)z^*$.

[7] 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.

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 account for potential curvature in the counterfactual density when estimating the earnings response from bunching.

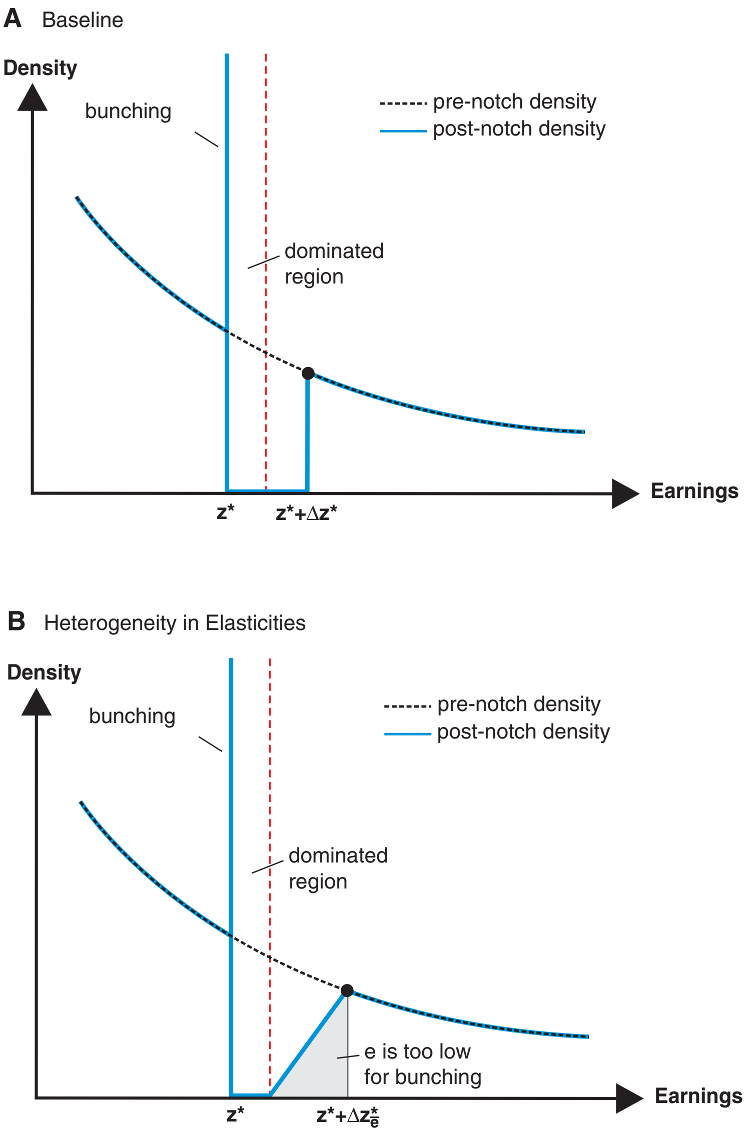
We now consider the following extensions of the model: heterogeneity in elasticities, optimization frictions, dynamics, and extensive responses. Figure II illustrates the effect of a notch on the density distribution in the benchmark model (Panel A) and in various more general models (Panels B–D). To 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 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 $(z^*, z^* + \Delta z^*)$.

*1. 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 $\left(z^*, z^* + \Delta z^*_e\right)$, where $\Delta z^*_e$ is increasing in $e$ and takes the value $\Delta z^D$ for $e = 0$. The post-notch earnings density in the full population is illustrated by the solid 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 gray 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 $E\left[\Delta z^*_e\right]$. 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

$$(8) \qquad B = \int_e \int_{z^*}^{z^* + \Delta z^*_e} \tilde{h}_0(z, e)dzde \approx h_0(z^*)E\left[\Delta z^*_e\right],$$

where the approximation again assumes that the counterfactual density is locally constant in earnings (but not elasticities). Using equation (8), estimates of excess bunching and the counterfactual


A Baseline

B Heterogeneity in Elasticities

FIGURE II

Density Distributions under Different Model Extensions

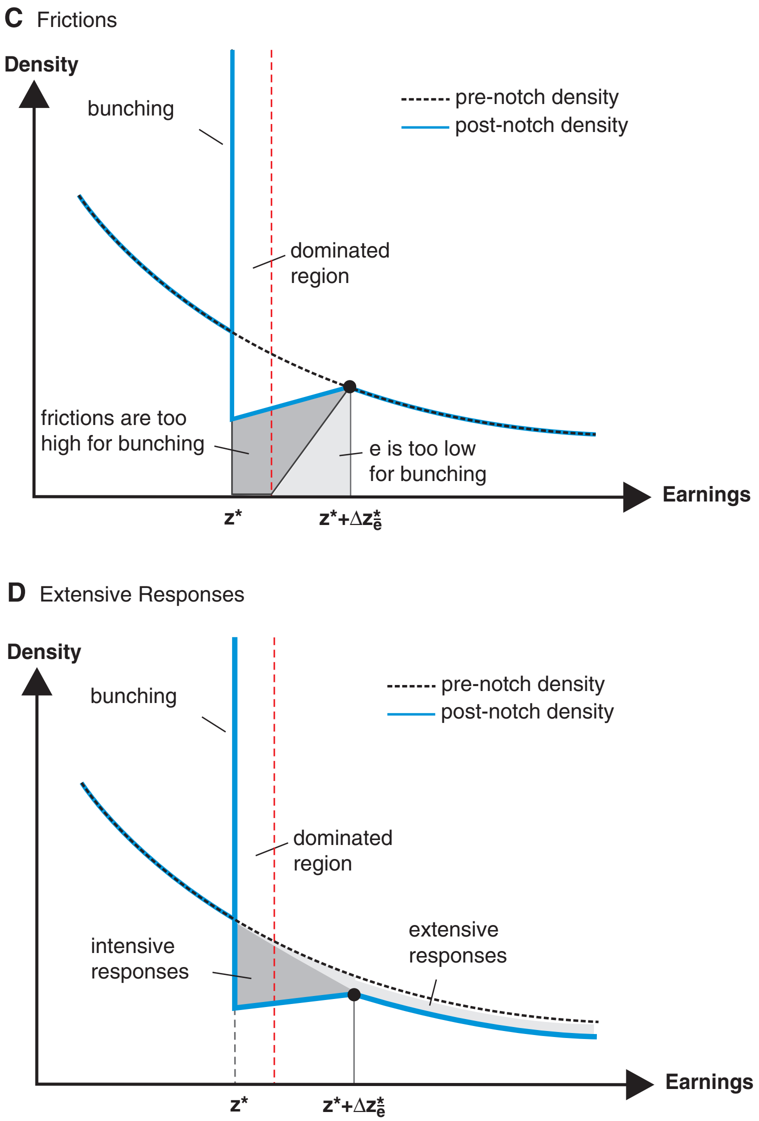


FIGURE II

Continued

earnings density reveal the average earnings response in the population.

*2. 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 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), whereas 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 $\left(z^*, z^* + \Delta z^*_{\bar{e}}\right)$ 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 gray area in the figure consists of those who do not respond because of low structural elasticities, and the dark gray 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 $\Delta z^*_e$ 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).[8] Although 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 that provide, respectively, lower and upper bounds on the structural elasticity.

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

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

$$B = \int_e \int_{z^*}^{z^*+\Delta z_e^*} (1-a(z,e))\, \tilde{h}_0\,(z,e)\, dzde \approx h_0\,(z^*)(1-a^*)\,E\left[\Delta z_e^*\right], \tag{9}$$

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 $z \in \left(z^*, z^* + \Delta z_e^*\right)$ and all $e$. In the foregoing expression, $E\left[\Delta z_e^*\right]$ is the average structural response not affected by frictions while $(1-a^*)E\left[\Delta z_e^*\right]$ 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 $a^* \equiv \int_{z^*}^{z^*+\Delta z^D} h(z)dz \Big/ \int_{z^*}^{z^*+\Delta z^D} h_0(z)dz$.

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 (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 (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 smooth, this effect makes $a(z,e)$ increasing on the bunching segment $\left(z^*, z^* + \Delta z^*_e\right)$.[9] 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 $\left(z^*, z^* + \Delta z^*_{\bar{e}}\right)$ can be explained by low elasticities in a frictionless world (light gray area in Panel C of Figure II). 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 II, 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

[9] 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.

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.2 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 decreasing above the notch and feature no hole.

*3. 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 multiperiod 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 $K > 1$.[10]

*4. 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 fixed cost of participation $q$ that is smoothly distributed in the population. Extending the formulation (1), utility from participation is given by $u(z - T(z), z) - q$ while utility from nonparticipation 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 nonparticipation. Denoting by $\bar{q}_0$ the threshold fixed cost under the baseline linear tax system,

[10] 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.

$T(z) = t \cdot z$, there will be extensive responses by those with $q \in (\bar{q}_0 - \Delta\bar{q}, \bar{q}_0)$ where

$$\Delta\bar{q} = u((z^* + \epsilon)(1-t),\ z^* + \epsilon) - u(z^*(1-t),\ z^*), \tag{10}$$

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.[11] 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.[12] These results imply that extensive responses affect the density distribution as illustrated in Panel D of Figure II.

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 particular, the bunching-hole method 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

[11] 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.

[12] 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.

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.

### *II.B. 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 $\Delta z^*$ 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

$$(11) \quad t^* \equiv \frac{T(z^* + \Delta z^*) - T(z^*)}{\Delta z^*} = t + \frac{\Delta t \cdot (z^* + \Delta z^*)}{\Delta z^*} \approx t + \frac{\Delta t \cdot z^*}{\Delta z^*},$$

where the approximation requires that $\Delta t$ 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

$$(12) \qquad e_R \equiv \frac{\Delta z^*/z^*}{\Delta t^*/(1-t^*)} \approx \frac{(\Delta z^*/z^*)^2}{\Delta t/(1-t)}.$$

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 III illustrates the relationship between the reduced-form and structural approaches using a budget set diagram. The reduced-form formula (12) treats the response to the notch $\Delta z^*$ as if it were generated by the kink shown by the intersection of the lower budget segment (slope $1 - t$) with the solid black line (slope $1 - t^*$). 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

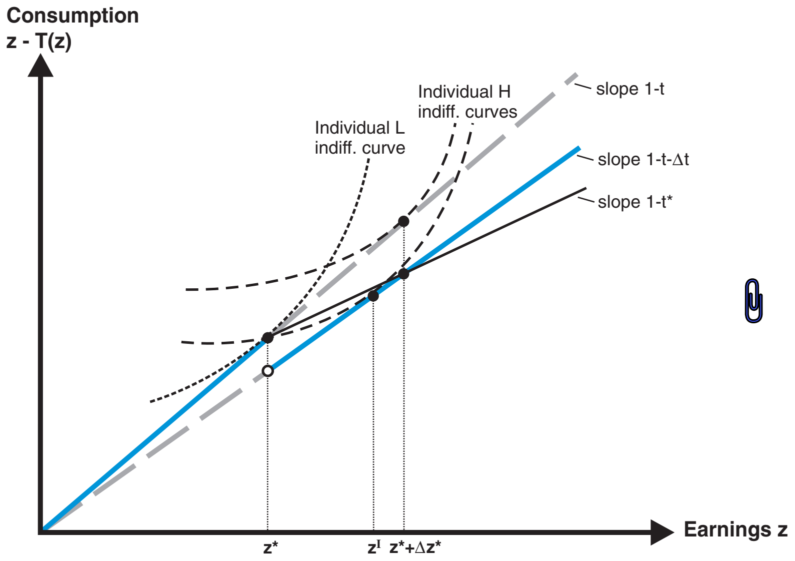


FIGURE III

Reduced-Form Approximation of Earnings Elasticity

not become a buncher if faced with this kink. In this case, the bunching response to the notch $\Delta z^*$ overstates the bunching response that would be created by the kink $\Delta t^*$, implying that the reduced-form elasticity $e_R$ constitutes an upper bound. The key reason 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.[13]

[13] 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.3 in the Online Appendix shows absolute and relative bias as a function of $\Delta z^*/z^*$, assuming that true preferences are quasi-linear as in (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 $e_R > 0$.

### II.C. 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).

*1. Standard Case.* Consider the (hypothetical) empirical density distribution in Panel A of Figure IV. 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 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

$$c_j = \sum_{i=0}^{p} \beta_i \cdot \left(z_j\right)^i + \sum_{i=z_L}^{z_U} \gamma_i \cdot \mathbf{1}\left[z_j = i\right] + \nu_j, \tag{13}$$

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 (13) omitting the contribution of the dummies in the excluded range, that is, $\hat{c}_j = \sum_{i=0}^{p} \hat{\beta}_i \cdot \left(z_j\right)^i$. Excess bunching and missing mass are estimated as the difference between the observed and counterfactual bin counts in the relevant earnings ranges, $\hat{B} = \sum_{j=z_L}^{z^*} \left(c_j - \hat{c}_j\right)$ and $\hat{M} = \sum_{j>z^*}^{z_U} \left(\hat{c}_j - c_j\right)$. 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 IV.

Standard errors are calculated using a bootstrap procedure in which we generate a large number of earnings distributions (and associated estimates of each variable) by random resampling of residuals in (13). The standard error of each variable is defined

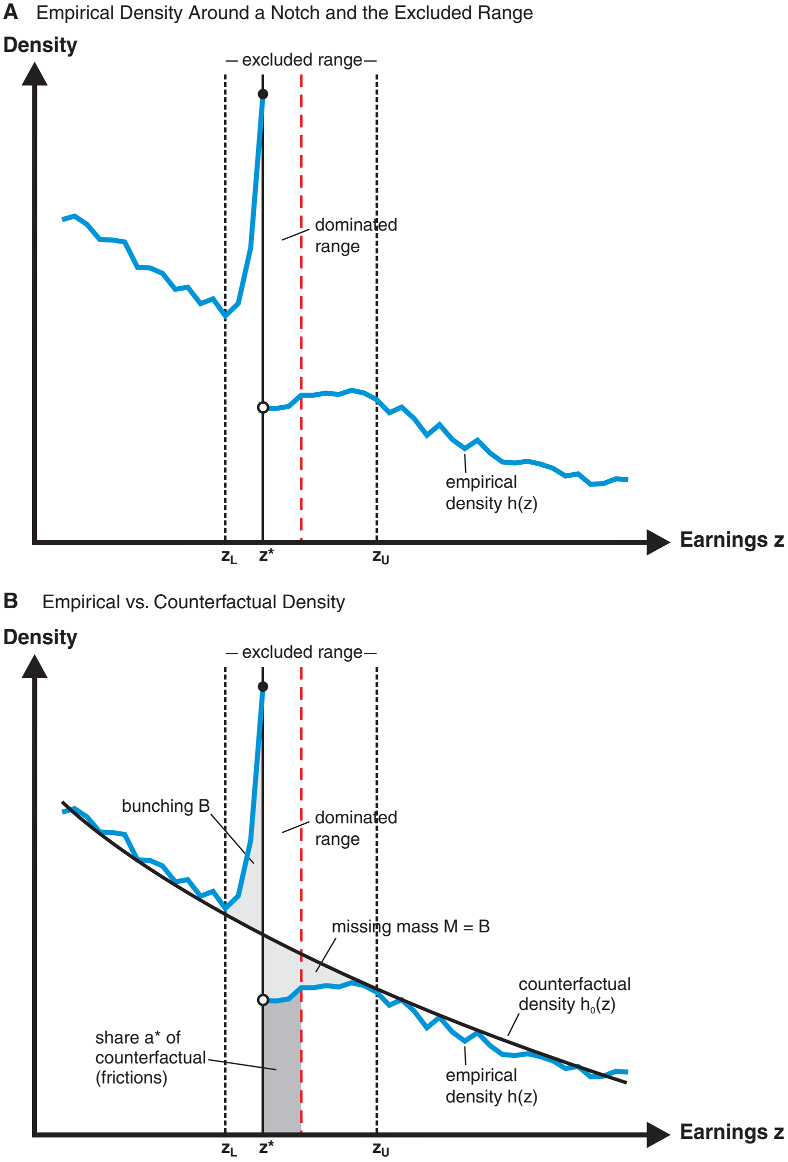


FIGURE IV

Estimating the Counterfactual Density from an Empirical Density as the standard deviation in the distribution of estimates of the given variable.

The approach relies on a credible determination of the excluded range $[z_L, z_U]$. Because 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, because 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, have $\hat{M}^0 \ll \hat{B}^0$), the upper bound is increased in small increments and the counterfactual reestimated every time until we achieve $\hat{M}^k = \hat{B}^k$. The resulting estimate $\hat{z}_U = \hat{z}_U^k$ not only represents 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.[14]

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 II (Panel D), intensive responses are concentrated in a triangular area close to the cutoff, whereas 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

[14] 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 I (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).

must be equal to bunching mass $B$. Because 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 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 $p$).

*2. 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. (1) The counterfactual distribution is smooth such that excess bunching $B$ identifies a behavioral response.[15] (2) Bunchers come from a continuous set $M = B$ above the cutoff such that there exists a well-defined marginal buncher. (3) 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 (1–2) are quite weak, assumption (3) 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 (1–2) along with a much stronger third assumption (3') that the

[15] 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).

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.

*3. 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, although 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, etc.) in both monthly and annual terms.

There are two conceptual points to note about round-number bunching. First, since notches are themselves located at salient round numbers, implementing the specification (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. To construct such round-number counterfactuals convincingly, we account for the underlying anatomy of rounding 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

$$c_j = \sum_{i=0}^{p} \beta_i \cdot \left(z_j\right)^i + \sum_{r \in R,\, 12 \cdot R} \rho_r \cdot \mathbf{1}\left[\frac{z_j}{r} \in \mathbb{N}\right] + \sum_{i=z_L}^{z_U} \gamma_i \cdot \mathbf{1}\left[z_j = i\right] + \nu_j, \tag{14}$$

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 (14) omitting the contribution of the dummies around the notch, but not omitting the contribution of round-number dummies.

## III. Application to Tax Notches in Pakistan

### III.A. *Income Tax and Enforcement System*

The personal income tax in Pakistan currently raises revenue of 1.1% of gross domestic product (GDP), or 11% of total tax revenue, and the share of registered taxpayers in the working-age population is less than 2%.[16] 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: (1) 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), (2) 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 nonregistrations. 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.

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 V shows the average tax rate as a function of taxable income in Pakistani rupees (PKR) for self-employed individuals (tax years 2006–2009) and wage earners (tax years 2006–2007).[17]

[16] See World Bank (2009).

[17] 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.

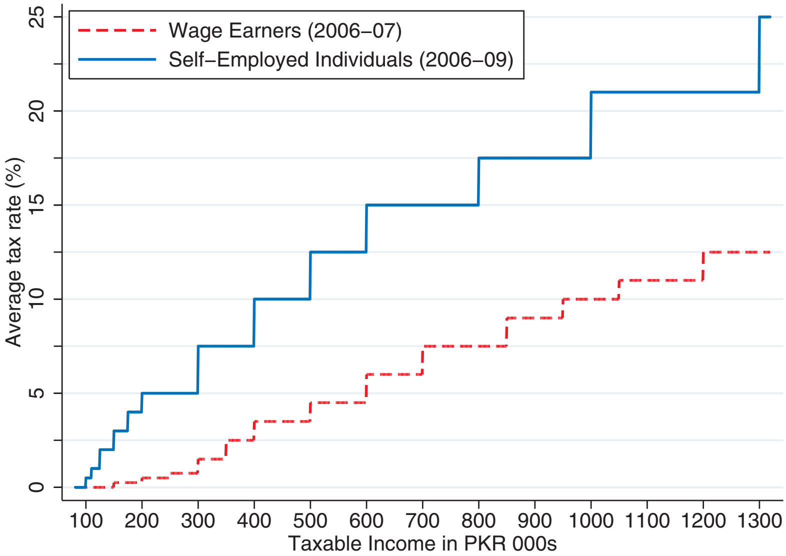


FIGURE V

Personal Income Tax Schedules in Pakistan

The figure shows the average tax rate as a function of annual taxable income for wage earners in 2006–7 (dashed line) and for self-employed individuals in 2006–9 (solid line). Taxable income is shown in thousands of Pakistani rupees (PKR), with the PKR-USD exchange rate varying from 60 to 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 lowtax side of the notch. The tax rate on self-employed individuals increases from 0 to 25% over 13 notches, while the tax rate on wage earners increases from 0 to 20% over 20 notches (the first 13 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.

We note the following about these schedules. First, the tax rate on self-employed individuals increases from 0 to 25% over 13 notches, while the tax rate on wage earners increases from 0 to 20% over 20 notches (the first 13 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.[18]

Registered taxpayers are required to file income tax returns unless they meet certain filing exemption requirements.[19] The tax return is shown in Figure A.4 of the Online Appendix.[20] 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

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

[19] In particular, wage earners are exempt from filing if (1) wage income is below 500K, (2) the employer has filed a tax return (third-party report), and (3) the taxpayer has no nonwage income. For such nonfilers, 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 article 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.

[20] 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 prevalidation 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.

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.

### *III.B. 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 have 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 noncompliance. 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.

### *III.C. Results for Self-Employed Individuals*

This section presents empirical results for self-employed males.[21] As explained, self-employed individuals have a tendency to report taxable income in round numbers, 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 (13).[22] 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 (14).

Figure VI presents evidence from the first 10 notches of the tax schedule for the non-rounder sample between 2006 and 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 solid 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, although 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

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

[22] 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.

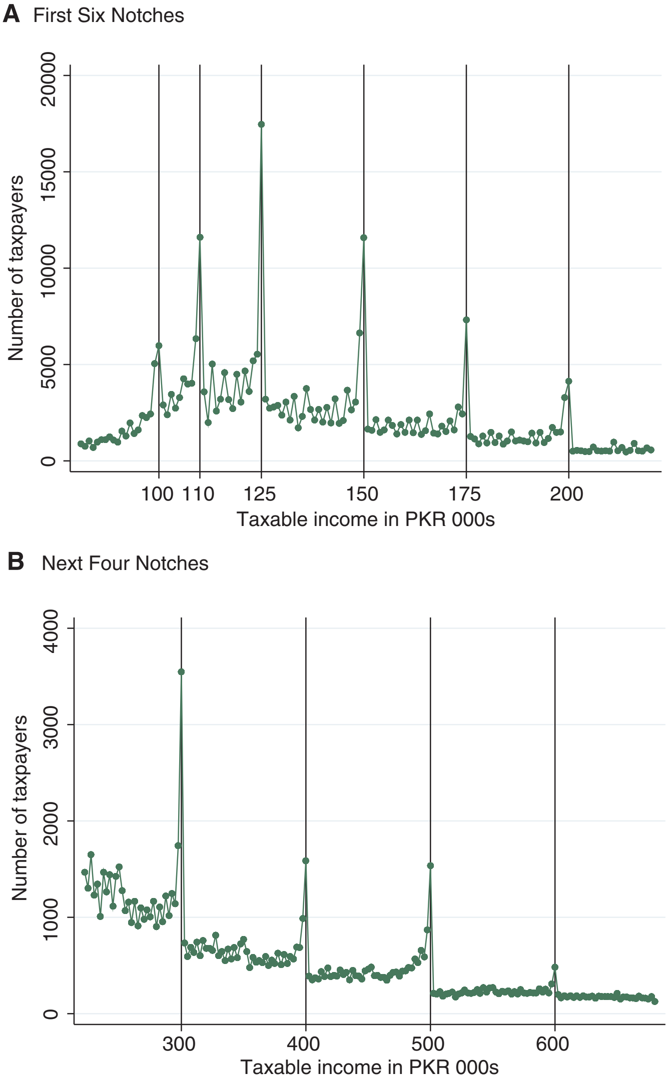


FIGURE VI

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

The figure shows the empirical distribution of taxable income (dotted graph) and the counterfactual distribution (solid graph) for self-employed individuals (non-rounder sample) from 2006 to 2009. 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 vertical solid lines, upper bounds of

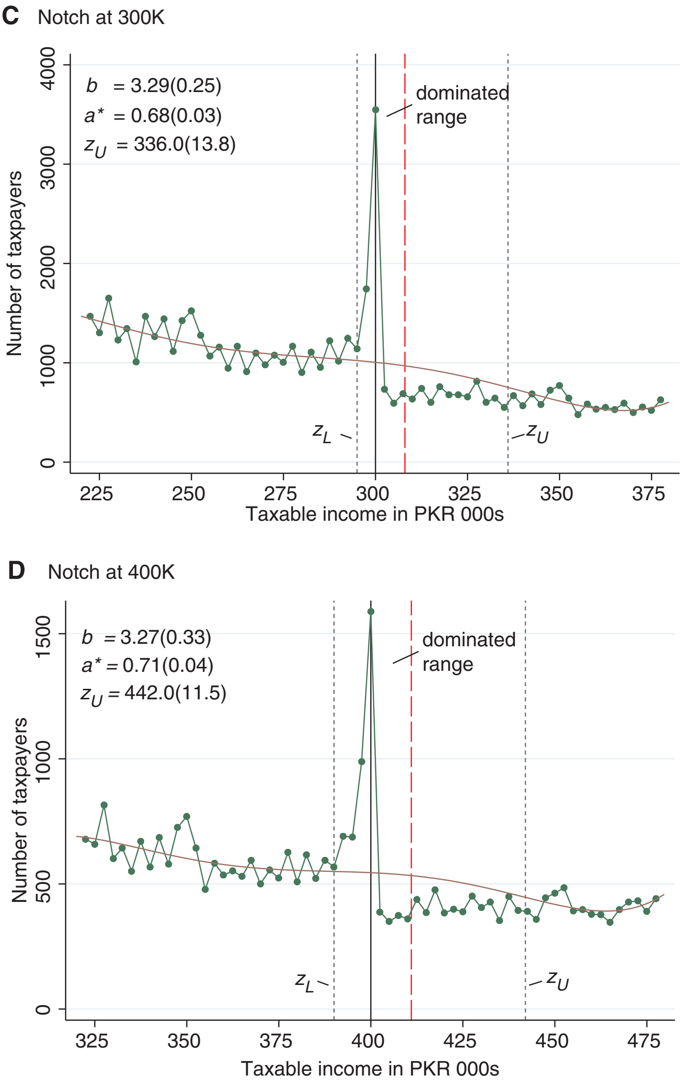
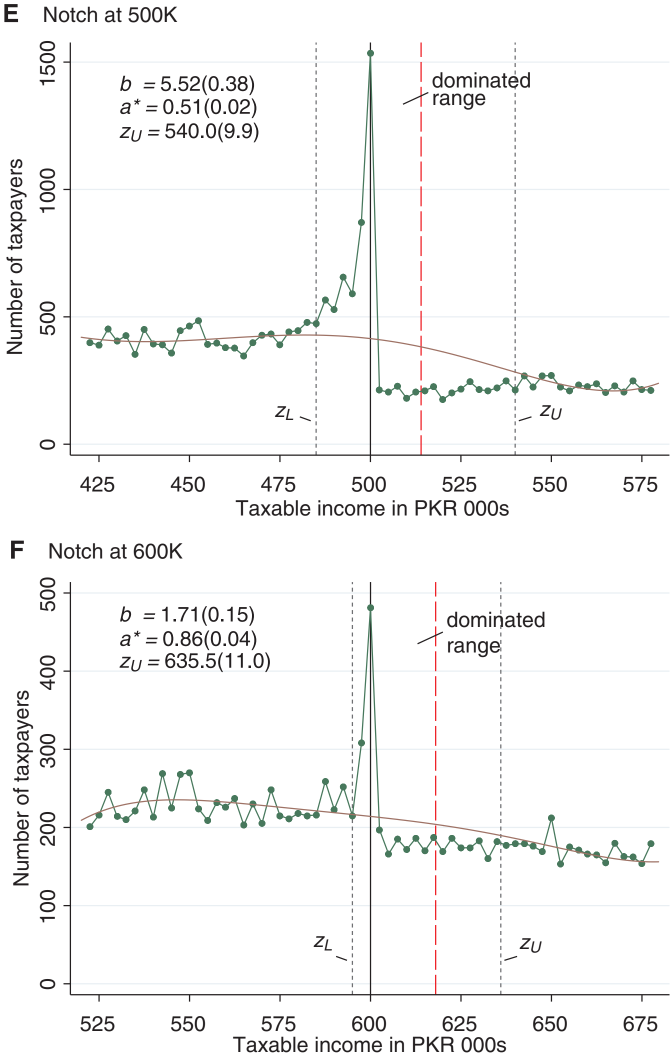


FIGURE VI

Continued

dominated regions are marked by vertical long-dashed lines, and excluded ranges $[z_L,\ z_U]$ are marked by vertical short-dashed 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.



E Notch at 500K

F Notch at 600K

FIGURE VI

Continued

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 II.C, 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 II. It could also include discrete intensive responses between the interiors of brackets, possibly driven by optimization frictions that prevent some individuals to target the region close to the notch point. Such effects cannot be ruled and our bunching approach cannot capture them. Hence, although 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 VI 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 (13).[23] The excluded range $[z_L, z_U]$ is demarcated by vertical short-dashed lines and the upper bound of the strictly dominated region is demarcated by a vertical long-dashed line.[24] Each panel shows

[23] Figure A.5 in the Online 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.

[24] 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 II.C) and the locally estimated counterfactual is supposed to capture what would

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 II, the amount of bunching absent frictions $b/(1-a^*)$ is two to seven 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 VII turns to the full sample and is constructed exactly as the preceding figure. 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 II. There is more rounding at the bottom of the distribution than at the top, consistent with the earlier remark that rounding is a

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.

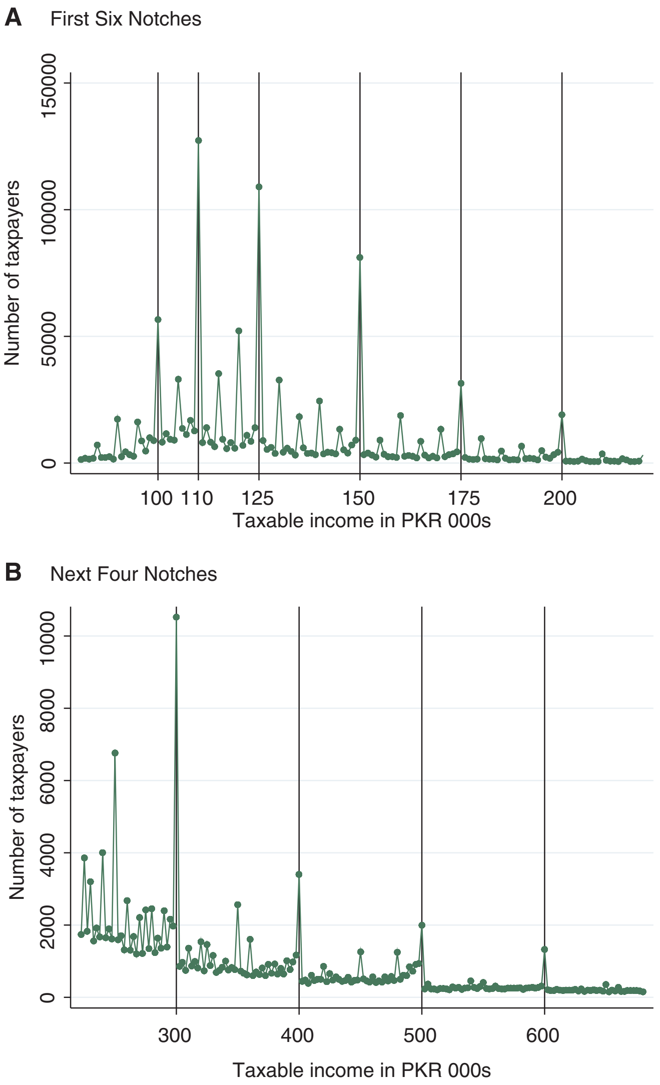


FIGURE VII

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

The figure shows the empirical distribution of taxable income (dotted graph) and the counterfactual distribution (solid graph) for self-employed individuals (full sample) from 2006 to 2009. 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 vertical solid lines, upper

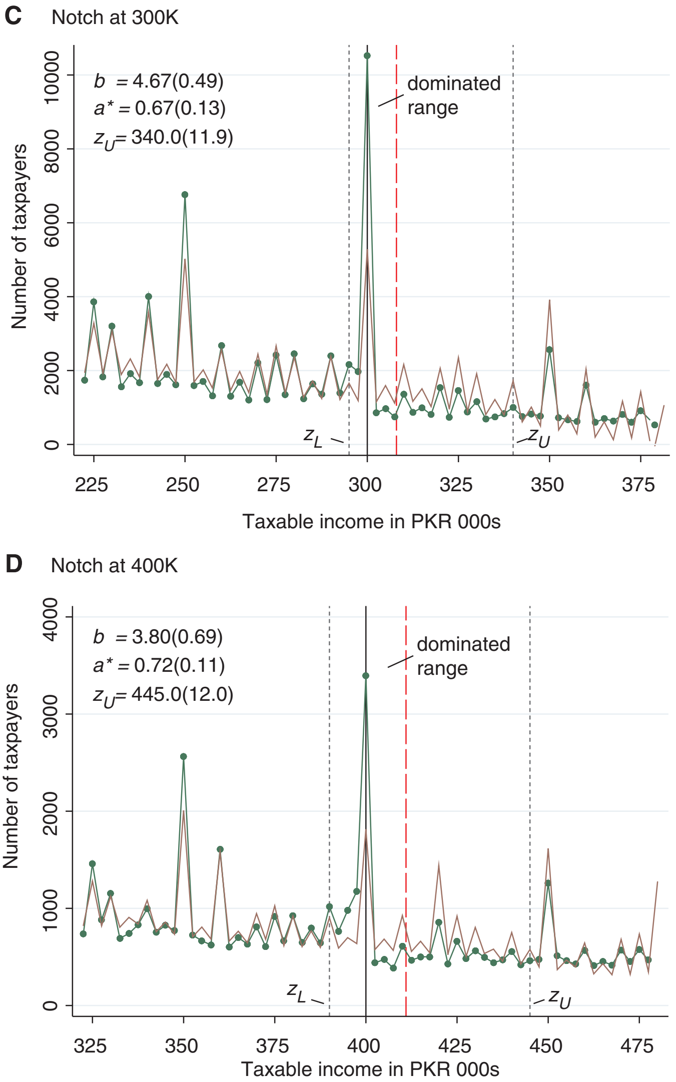
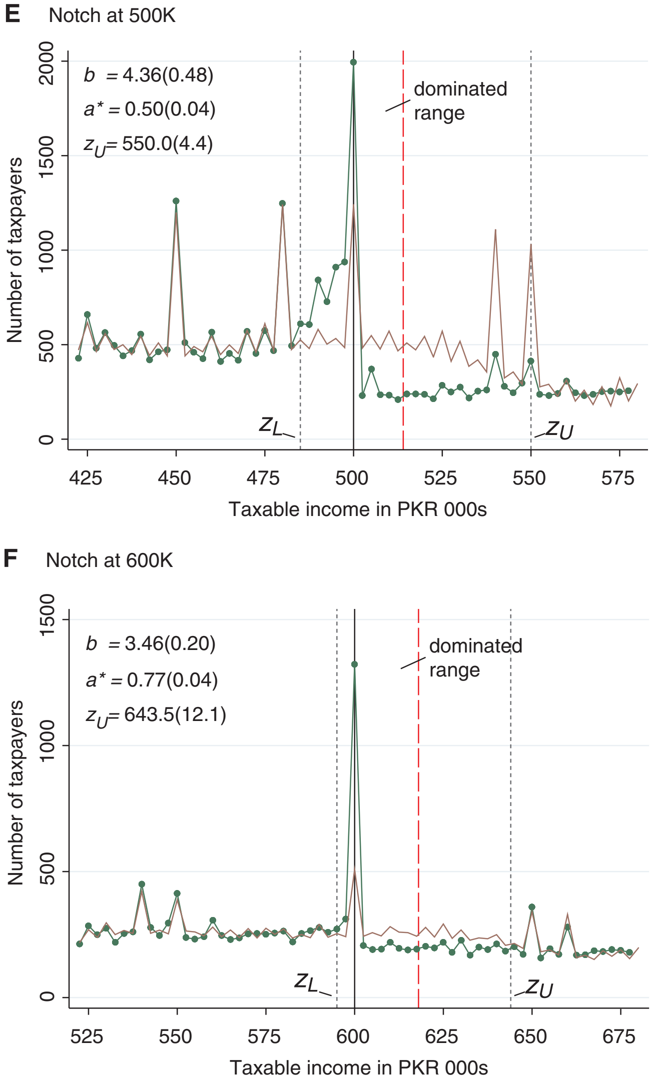


FIGURE VII

Continued

bounds of dominated regions are marked by vertical long-dashed lines, and excluded ranges [$z_L$, $z_U$] are marked by vertical short-dashed 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.



E Notch at 500K

F Notch at 600K

FIGURE VII

Continued

side effect of poor record keeping. The counterfactual distribution is estimated as a fifth-order polynomial with round-number fixed effects as specified in (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 and 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.[25]

Taking advantage of the longitudinal aspect of the data, Table I 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 III.A).[26]

[25] Figure A.6 in the Online Appendix considers lower and higher polynomial degrees for the full sample.

[26] 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 III.A). This system works as follows. Taxpayers who underestimate and underpay their income taxes (given their self-assessed taxable income) are labeled 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.

TABLE I

DYNAMICS OF DOMINATED, BUNCHING, AND INCONSISTENT BEHAVIOR

| | | Dominated Behavior | | | | Bunching Behavior | | | | Inconsistent Reporting | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | #Obs. | | | | | | | | | | | |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) |
| | | Total | 2-Year | 3-Year | 4-Year | Total | 2-Year | 3-Year | 4-Year | Total | Among Dominated | Among Bunchers |
| **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) |

TABLE I

(CONTINUED)

| Year | #Obs. | Dominated Behavior | | | | Bunching Behavior | | | | Inconsistent Reporting | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) |
| | | Total | 2-Year | 3-Year | 4-Year | Total | 2-Year | 3-Year | 4-Year | Total | Among Dominated | Among Bunchers |
| **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 to 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 13 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.

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 had moved out of such 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.7 in the Online Appendix shows graphically that excess bunching becomes stronger and dominated behavior slightly weaker over the sample period, consistent with the findings in Table I. 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 nonparametric evidence above with the conceptual framework in Section II. Such elasticities can be obtained by estimating the earnings response of the marginal buncher and applying the parametric relationship (5) or the reduced-form approximation (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 II, which shows the notch point in column (1), the average tax rate jump in column (2),

TABLE II

STRUCTURAL EARNINGS ELASTICITIES FOR SELF-EMPLOYED INDIVIDUALS (NON-ROUNDER SAMPLE)

| | | | Frictions $a^*$ | | Earnings Response $\Delta z^*$ | | Structural Elasticity e | | Reduced-Form Elasticity $e_R$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| (1)<br>Notch<br>Point | (2)<br>ATR<br>Jump $\Delta t$ | (3)<br>Dominated<br>Range $\Delta z^D$ | (4)<br>Full<br>Range $\Delta z^D$ | (5)<br>Lower<br>Range $\Delta z^D/2$ | (6)<br>Bunching-Hole<br>Method | (7)<br>Conv.<br>Method | (8)<br>Bunching-Hole<br>Method | (9)<br>Conv.<br>Method | (10)<br>Bunching-Hole<br>Method | (11)<br>Conv.<br>Method |
| 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 to 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 $1/(1-a^*)$ 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.

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).[27]

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 II. 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. Although 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.

[27] In the calculation of standard errors (using the bootstrap method), 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.

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, Slemrod, and Giertz 2012). Third, because 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.

### *III.D. 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% 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 20 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 VIII presents nonparametric evidence on behavioral responses and frictions for wage earners in 2006–2007, and is constructed exactly as the analogous figures in the

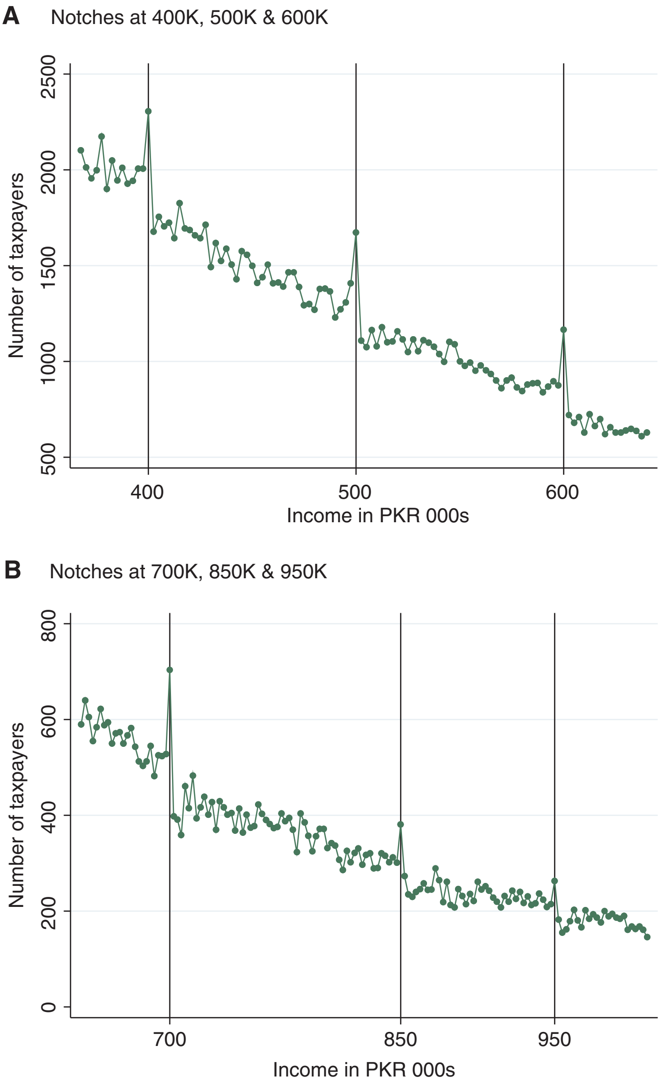
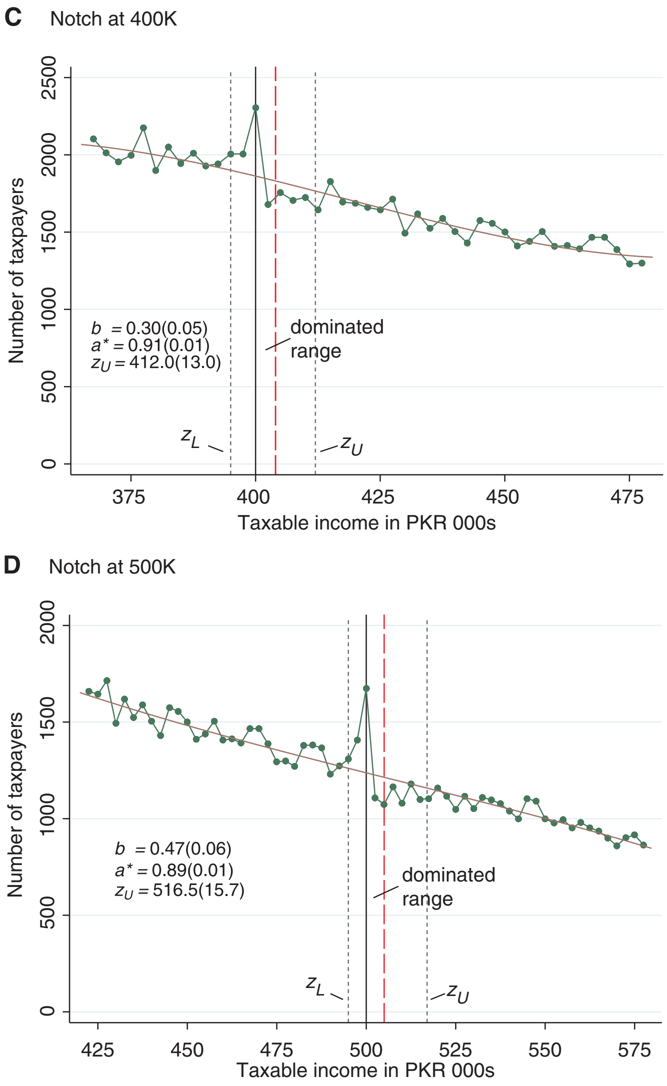


FIGURE VIII

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

The figure shows the empirical distribution of taxable income (dotted graph) and the counterfactual distribution (solid graph) for wage earners (non-rounder sample) from 2006 to 2007. 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 vertical solid lines, upper bounds of dominated regions are marked by vertical



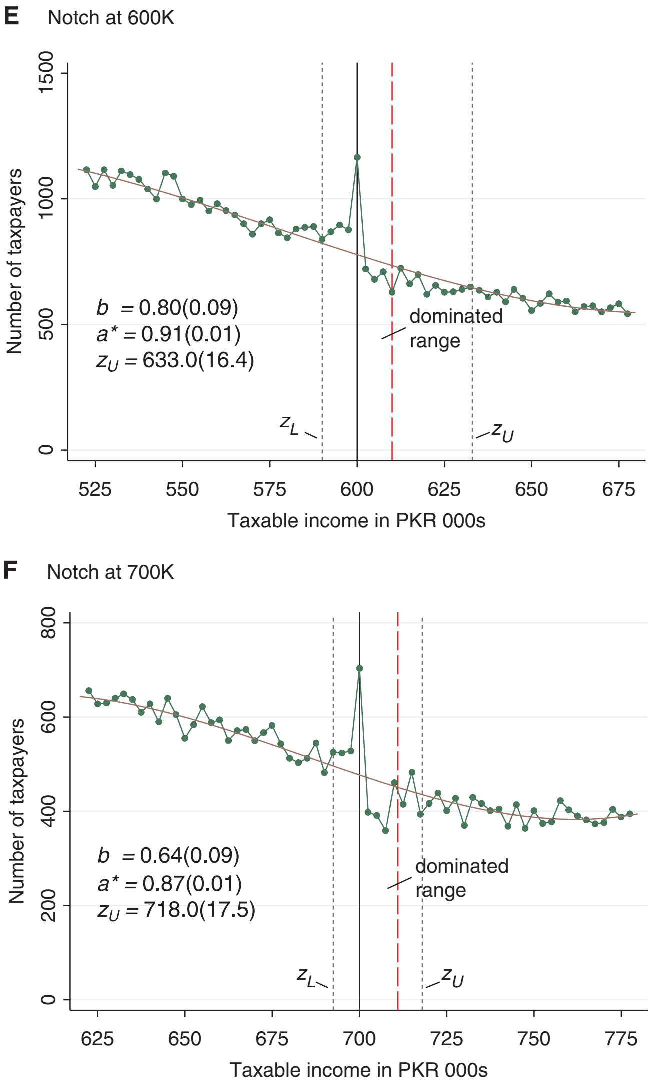
C Notch at 400K

D Notch at 500K

FIGURE VIII

Continued

long-dashed lines, and excluded ranges $[z_L, z_U]$ are marked by vertical short-dashed 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.



E Notch at 600K

F Notch at 700K

Figure VIII

Continued

previous section.[28] The estimation of the counterfactual distribution is based on a third-order polynomial (instead of a fifth-order polynomial) as the distribution for wage earners has less curvature than the distribution for self-employed individuals.

The main findings in Figure VIII 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.2 (Panel C) in the Online 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 III 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

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

TABLE III

STRUCTURAL EARNINGS ELASTICITIES FOR WAGE EARNERS (NON-ROUNDER SAMPLE)

|  |  |  | Frictions a* |  | Earnings Response Δz* |  | Structural Elasticity e |  | Reduced-Form Elasticity $e_R$ |  |
|---|---|---|---|---|---|---|---|---|---|---|
| (1)<br>Notch<br>Point | (2)<br>ATR<br>Jump Δt | (3)<br>Dominated<br>Range $\Delta z^D$ | (4)<br>Full<br>Range $\Delta z^D$ | (5)<br>Lower<br>Range $\Delta z^D/2$ | (6)<br>Bunching-Hole<br>Method | (7)<br>Conv.<br>Method | (8)<br>Bunching-Hole<br>Method | (9)<br>Conv.<br>Method | (10)<br>Bunching-Hole<br>Method | (11)<br>Conv.<br>Method |
| 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 to 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*) 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.

locally constant frictions. Second, earnings responses are mostly between 2% and 5% 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, generally around 0.05 or lower.

### *III.E. 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 IX presents nonparametric 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: (1) 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

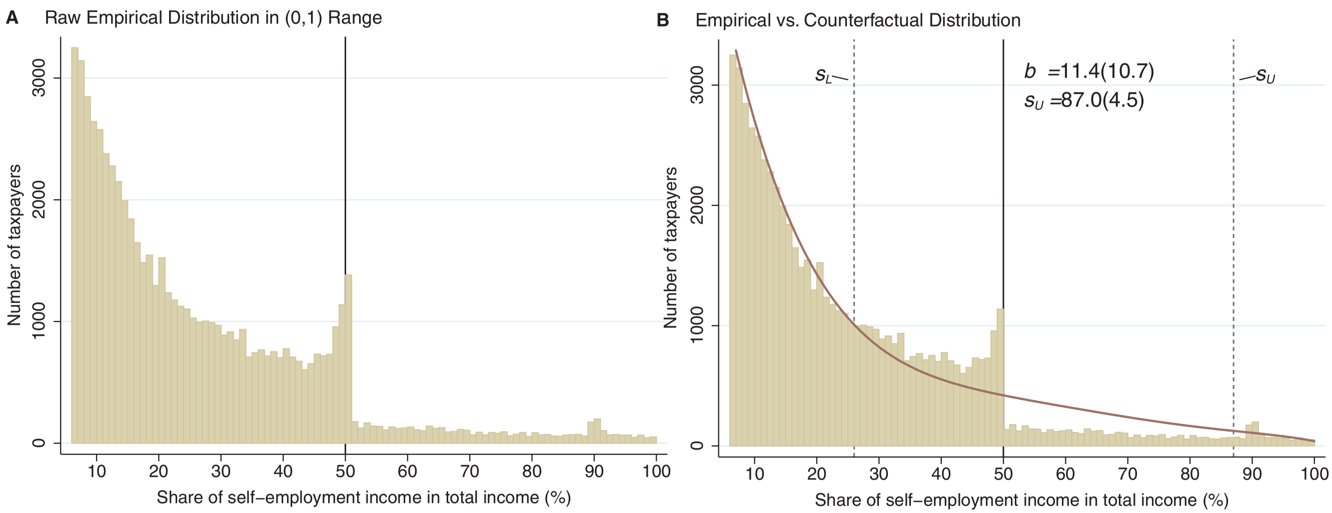


FIGURE IX

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

The figure shows the empirical distribution of the self-employment income share (in bars) and the counterfactual distribution (solid graph) for individuals with both self-employment income and wage income from 2006 to 2009. Panel A includes all observations with a self-employment income share between 0 and 1, and Panel B drops observations precisely at the cutoff value of 50%. The counterfactual is estimated by fitting a fifthorder polynomial to the empirical distribution, excluding data around the notch, as specified in equation (13). The notch point is marked by a vertical solid line and the excluded range [$sL$, $sU$] is marked by vertical short-dashed 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 $sU$ has been estimated to ensure that missing mass equals bunching mass. Standard errors are shown in parentheses.

self-employment and wage income, or (2) 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%, whereas 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 percentage 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 IX, considerably larger than the notches considered earlier.[29]

[29] 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.

## IV. 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 nonparametrically 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 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.

London School of Economics and CEPR
London School of Economics

## Supplementary Material

An Online Appendix for this article can be found at QJE online (qje.oxfordjournals.org).

## References

Besley, Timothy, and Torsten Persson, "Taxation and Development," in *Handbook of Public Economics*, Volume 5, Alan Auerbach, Raj Chetty, Martin Feldstein, and Emmanuel Saez, eds (Amsterdam: Elsevier, 2012).

Blundell, Richard, and Hilary W. Hoynes, "Has 'In-Work' Benefit Reform Helped the Labor Market?," in *Seeking a Premier Economy: The Economic Effects of British Economic Reforms, 1980–2000*, Richard Blundell, David Card, and Freeman, Richard B., eds (Chicago: University of Chicago Press 2004).

Blundell, Richard, and Andrew Shephard, "Employment, Hours of Work and the Optimal Taxation of Low Income Families," *Review of Economic Studies*, 79 (2012), 481–510.

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

Chetty, Raj, and Emmanuel Saez, "Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients," *American Economic Journal: Applied Economics*, 5, no. 1 (2013), 1–31.

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

Gruber, Jonathan, and David A. Wise, *Social Security and Retirement around the World* (Cambridge, MA: National Bureau of Economic Research, 1999).

Kleven, Henrik J., Martin B. Knudsen, Claus T. Kreiner, Søren Pedersen, and Emmanuel Saez, "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark," *Econometrica*, 79 (2011), 651–692.

Manoli, Day, and Andrea Weber, "Nonparametric Evidence on the Effects of Financial Incentives on Retirement Decisions (NBER Working Paper No. w17320, 2011).

Ramnath, Shanthi P., "Taxpayers' Response to Notches: Evidence from the Saver's Credit (University of Michigan Working Paper, 2009).

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

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

Sallee, James M., and Joel Slemrod, "Car Notches: Strategic Automaker Responses to Fuel Economy Policy," *Journal of Public Economics*, 96 (2012), 981–999.

Slemrod, Joel, "Buenas Notches: Lines and Notches in Tax System Design (University of Michigan Working Paper, 2010).

World Bank, *Pakistan Tax Policy Report: Tapping Tax Bases for Development*. vol. 2 (Washington, DC: World Bank, 2009).

Yelowitz, Aaron S., "The Medicaid Notch, Labor Supply, and Welfare Participation: Evidence from Eligibility Expansions," *Quarterly Journal of Economics*, 110 (1995), 909–939