Diffuse Bunching with Frictions: Theory and Estimation*

Santosh Anagol Allan Davids
Benjamin B. Lockwood Tarun Ramadorai†

June 7, 2024

Abstract

We incorporate a general model of frictions into the bunching-based elasticity estimator. This model relies on fewer parameters than the conventional approach, replacing bunching window bounds with a single "lumpiness parameter," while matching rich observed bunching patterns such as sharp-peaked diffusion around tax kinks and depressed density in the dominated region above a notch. Simulations suggest that in the presence of frictions, conventional methods may underestimate elasticities with overstated confidence. Our method draws information from the spread of bunching mass around kinks and asymmetry around notches, revealing the size of frictions, unobserved costs, and kink vs. notch misperceptions. Estimating this model on South African administrative tax data, we find that individuals and firms appear to treat the bottom zero-to-positive tax kink like a notch, and we uncover differences in lumpiness between wage earners vs. the self-employed and between firms with vs. without paid tax practitioners. *JEL* codes: H30, J20, O12

*We wish to acknowledge the National Treasury of South Africa for providing us with access to anonymized tax administrative data. We thank Analytics at Wharton and the Penn Wharton Budget Model for funding support. The views expressed in this paper are our own and do not necessarily reflect the views of the National Treasury of South Africa. We are grateful to Wian Boonzaaier, Ana Gamarra Rondinel, Henrik Kleven, Dylan Moore, Jacob Mortenson, Alex Rees-Jones, Joel Slemrod, Jakob Søgaard, David Thesmar, Andrew Whitten, Eric Zwick, and seminar participants at the University of Pretoria, Economic Research South Africa (ERSA), the European Bank for Reconstruction and Development, Imperial College, IIPF 2023, LAGV 2021, LMU Munich, CREST, the NBER Public Economics Meetings, the Toulouse School of Economics, the University of Cape Town, and the South African Revenue Services for helpful comments and to Michael Partridge, Afras Sial, and Laila Voss for excellent research assistance. All errors are our own. This paper subsumes and replaces the working paper titled "Do Firms Have a Preference for Paying Exactly Zero Tax?"

†Anagol: Wharton School, University of Pennsylvania. Email: anagol@wharton.upenn.edu. Davids: School of Economics, University of Cape Town. Email: allan.davids@uct.ac.za. Lockwood: Wharton School, University of Pennsylvania and NBER. Email: ben.lockwood@wharton.upenn.edu. Ramadorai: Imperial College London and CEPR. Email: t.ramadorai@imperial.ac.uk

1 Introduction

The elasticity of taxable income is among the most central parameters in public economics, appearing as an input in many economic forecasts. It is a key statistic in models of optimal taxation, governing the optimal asymptotic top tax rate on high earners as well as the revenue-maximizing tax rate. One influential estimation strategy, proposed by Saez (2010), seeks to quantify this elasticity by measuring the amount of excess bunching mass in the income distribution around tax bracket thresholds where there is a change in the marginal tax rate (a "kink") or tax level (a "notch").[1]

The conventional bunching estimator takes as its input the excess mass in an income density around a tax kink or notch relative to a "counterfactual density" that would arise if the tax were linear. This excess mass is converted into an elasticity using a bunching estimator formula, which Saez (2010) derives using a frictionless model that predicts an atom of excess mass at the kink. In contrast with this prediction, the bunching observed in empirical distributions is typically diffuse, obscuring the distinction between bunching induced by tax incentives and fluctuations in the underlying productivity distribution that would appear even under a linear tax. A solution proposed by Chetty et al. (2011), which has become the conventional approach, is to measure diffuse excess bunching mass as the difference between the observed density and a smooth function fitted to the observed density outside of a visually specified "bunching window" around the kink. Kleven and Waseem (2013) extend this approach to notches.[2] By abstracting from the source of the frictions that produce diffusion, these approaches leave open the question of whether they successfully recover the structural elasticity of taxable income in the presence of those frictions. Moreover, by reducing the pattern of density distortions around the tax bracket threshold to a single number representing excess mass, they potentially discard other useful information about taxpayers' behavioral responses to the kink or notch.

In this paper, we relax the assumption that agents can perfectly target income when faced with kinks and notches in the tax schedule. We present a tractable model of income frictions that can match key features of empirical behavior around tax kinks and notches. It is also parsimonious, reducing the total number of model parameters by introducing a single, informative money-metric parameter governing income frictions while eliminating the need

[1] For examples, see Chetty et al. (2011), Kleven and Waseem (2013), Mortenson and Whitten (2020), Rees-Jones (2018) and others reviewed in Kleven (2016). Bunching estimation has also been applied to domains of retirement incentives (Manoli and Weber, 2016), mobile phone services (Grubb and Osborne, 2015), and educational test scores for both students (Diamond and Persson, 2016; Dee et al., 2019) and teachers (Brehm, Imberman and Lovenheim, 2017), marathon times (Allen et al., 2017), and home sales (Andersen et al., 2022).

[2] To account for non-zero density at "dominated incomes" above the notch, Kleven and Waseem (2013) assume a fraction of taxpayers are unresponsive, leaving unmodeled the source of diffusion in (one-sided) excess mass.

to specify bounds for a bunching window, either visually or algorithmically.

Concretely, we consider the class of models with "sparsity-based frictions," in which agents choose between discrete income opportunities drawn from around the target income that they would select in the absence of frictions. Sparsity-based frictions can arise from a range of microfoundations including search or adjustment costs (Chetty et al., 2011; Chetty, 2012; Kleven and Waseem, 2013; Gelber, Jones and Sacks, 2020; Mavrokonstantis and Seibold, 2022), lumpy adjustment (Rees-Jones, 2018), unpredictable bargaining outcomes (Andersen et al., 2022), and inattention (Sims, 2003; Gabaix, 2014).[3] We show that models which generate sparsity-based frictions share a common structure and are well approximated by a limiting case—in a sense formalized in our Proposition 3—of "uniform sparsity," in which income opportunities are drawn from a Poisson process with a single "lumpiness parameter" governing the average distance between income opportunities.

The uniform sparsity model predicts rich features of empirical bunching behavior around tax kinks and notches. Tax kinks induce "tent-shaped" (i.e., sharp-peaked and fat-tailed) symmetric bunching, as the many taxpayers who would exactly bunch under the frictionless model select their income opportunity nearest to the kink. This is consistent with empirically observed patterns (cf. Saez, 2010; Mortenson and Whitten, 2020).

Tax notches, in contrast, produce asymmetric bunching in the model, with diffuse excess mass below the notch and depressed, yet still positive, density at "dominated incomes" just above the notch. Intuitively, although such incomes are dominated by earning at the threshold—where effort is lower and post-tax income is higher—some taxpayers' next-best opportunity is sufficiently far away that an opportunity in the dominated region is preferred. These patterns are also consistent with empirical evidence on notches, such as in Kleven and Waseem (2013).

We use the model to assess the performance of the conventional bunching estimator when agents face sparsity-based frictions. We first demonstrate that the conventional approach is valid in this setting, but only under specific conditions. If both the below- and above-kink counterfactual densities that would arise under each region's linear tax schedules are known, then applying the usual bunching estimator formula to the integral of diffuse excess mass relative to these counterfactual densities recovers the true structural elasticity of taxable income. However, commonly-applied estimation procedures that involve recovering unknown counterfactual densities by fitting a polynomial to the observed density outside of a bunching window can produce biased elasticity estimates in the presence of sparsity-based frictions. This is a general challenge for the conventional approach, regardless of the model of frictions.

[3] See Søgaard (2019) for a review of different models of frictions in the context of labor supply adjustments and bunching patterns.

On simulated data generated with sparsity-based frictions, the conventional approach underestimates the true elasticity by as much as 50 percent in some specifications, and it produces 95 percent confidence intervals that exclude the true parameter in over 90 percent of simulations. This mismeasurement arises from the difficulty of estimating counterfactual densities using excluded bunching windows. In the presence of sparsity-based frictions some bunching mass spills beyond the specified bunching window, which pulls upward the estimated counterfactual density in the region around a kink, in turn leading to an underestimate of the excess bunching mass and thus the elasticity.[4] By structurally accounting for the predicted shape of diffuse bunching, our proposed approach recovers the true elasticity in the presence of sparsity-based frictions.

We apply this model to administrative tax data on individuals and small businesses in South Africa. Several features of the South African context make it a natural setting for our analysis. The light data requirements of the bunching estimator make it particularly well-suited to emerging markets, where historical longitudinal data and quasiexperimental tax variation may be scarce. Moreover, motivated by the Feldstein (1999) notion of the elasticity of taxable income as a sufficient statistic for welfare analyses and deadweight loss, this estimation strategy recovers the full elasticity including forces such as reporting responses and evasion, which may be particularly relevant in developing countries. Finally, the income tax schedule in South Africa is piecewise-linear with three prominent tax kinks for small businesses (and more for individuals) around which taxpayers exhibit pronounced bunching behavior. Among small businesses, we estimate elasticities in the range of 0.2 to 0.5 at the middle and upper tax kinks, and in excess of 1 at the lowest kink. In contrast, the conventional approach estimates elasticities to be about one half as large, consistent with the mismeasurement we document in our simulations.

In addition to improving elasticity estimates in the presence of frictions, our approach allows us to recover other information that is discarded by conventional bunching estimation procedures. First, we can recover information about the quantitative magnitude of the frictions that produce bunching using the spread of the bunching mass. Our model predicts that for a given excess mass, and thus a given elasticity, the spread identifies the average distance—measured in dollars—between income opportunities in the uniform sparsity model. We find that small businesses with paid tax practitioners exhibit significantly less bunching diffusion than other firms, suggesting that they target their incomes more precisely in response to tax incentives. This heterogeneity is not detected by the conventional bunching estimator, which produces statistically indistinguishable elasticities for the two groups.

[4] This issue is distinct from the issue of short-run vs. long-run (or "micro" vs. "macro") elasticities, both of which are confounded by this mismeasurement of the counterfactual density.

Second, we can use the prediction that notches create *asymmetry* in the bunching mass to gain additional insights about taxpayer behavior. Illustrating this point, both small businesses and individuals exhibit a pattern of bunching around their bottom tax kink that looks distinctly like the behavior expected around a notch, with diffuse mass below the bracket threshold and a drop in the density just above it. Such "notch-like" behavior is consistent with taxpayers perceiving a loss from earning just above the threshold, e.g., due to a hassle cost of remitting a tax payment or a misconception that the higher marginal rate applies to inframarginal income below the threshold.

An advantage of our approach is that we can estimate the (real or perceived) money-metric "notch value" consistent with the degree of asymmetry in bunching behavior. We once again find heterogeneity across firms, with firms that use paid tax practitioners exhibiting less notch-like behavior, consistent with lower hassle costs of tax payments or a clearer understanding of the difference between average and marginal rates. Such insights are unavailable from the conventional bunching estimator, which requires knowing the notch value as an input in order to identify the size of the dominated income region and the elasticity.

Although our application focuses on bunching in response to the income tax, the approach can readily be applied in other settings with kinked or notched budget sets. Many bunching applications study responses at thresholds for tax instruments which combine statutory changes in marginal or discrete tax incentives with unknown—but potentially important—changes in behavioral frictions or compliance costs. An example is value-added-tax (VAT) exemption thresholds (as studied in Velayudhan, 2018; Liu and Lockwood, 2015), which typically have both a known increase in the tax rate and an unknown change in compliance costs. Estimating the elasticity with respect to the VAT rate using conventional bunching methods requires the researcher to assume a compliance cost and then use the residual bunching to estimate the elasticity. In contrast, our method allows both the elasticity and the revealed-preference compliance cost to be estimated based on observed bunching behavior.

Our approach to frictions is distinct from adjustment cost models, such as Gelber, Jones and Sacks (2020), in which agents either pay an adjustment cost and fully reoptimize or else do not respond at all. Such models produce predictions about bunching dynamics, in which bunching increases over time as more agents adjust. Our static model serves as a complement to such models, predicting diffusion even in the static (steady state) equilibrium.

In addition to building on the theoretical and empirical literature on bunching estimators and frictions, this paper relates to two other bodies of literature. First, our findings documenting "notch-like" behavior at statutory kink points contributes to the literature on behavioral frictions and misperceptions about the tax code. Rees-Jones (2018) uses

bunching behavior around the threshold at which taxpayers face a net refund or balance due in order to quantify their degree of loss aversion. Rees-Jones and Taubinsky (2020) experimentally study misperceptions of the income tax code, finding that a substantial share of respondents "irons," misinterpreting an average tax rate as the relevant marginal rate. Using exogenous variation in worker knowledge about a notch in the Norwegian income tax system, Kostøl and Myhre (2021) estimate that at least 30 percent of estimated optimization frictions are due to workers' imperfect knowledge about the tax system. Outside the domain of taxes, Ito (2014) presents evidence that consumers respond (at the margin) to average rather than marginal electricity prices.

Second, our empirical application contributes to the large literature quantifying behavioral responses to taxation in developing economies. Particularly relevant is the subset of papers estimating the elasticity of corporate taxable income (e.g., Devereux, Liu and Loretz, 2014), which is a particularly important parameter in emerging market economies, given their greater relative reliance on the corporate income tax base (Gordon and Li, 2009). For examples, see Best et al. (2015) in Pakistan, Bachas and Soto (2021) in Costa Rica, and Pillay (2021), Kemp (2019), Boonzaaier et al. (2019) and Lediga, Riedel and Strohmaier (2019) in our setting of South Africa.

The rest of the paper proceeds as follows. In Section 2, we show how sparsity-based frictions modify the frictionless model of bunching at tax kinks and notches. We then show that a wide range of models with sparsity-based frictions can be approximated by a limiting case of uniform sparsity in which income opportunities are drawn from a Poisson process, and which can be estimated using tractable computational methods. Section 3 compares the performance of our estimation method with conventional bunching estimators using simulated data with known underlying parameters. Section 4 presents our empirical application to South African tax data. Section 5 concludes.

2 Model

2.1 Baseline bunching model with frictionless choice

Our starting point is the canonical "bunching estimator" presented in Saez (2010), and illustrated in Figure 1. Taxpayers with heterogeneous productivities—for example, individuals earning labor income or firms earning profits—choose incomes z under an income tax $T(z)$. The top panel of Figure 1 plots after-tax income as a function of pre-tax income; for consistency with the bunching literature, we will refer to these as "consumption" (c) and "earnings" (z), respectively.

Different tax schedules cause taxpayers to select different incomes, giving rise to different

income distributions. In the example plotted in Figure 1, the linear tax $T_0(z)$ results in the cumulative distribution function (CDF) of incomes labeled $H_0(z)$, while the linear tax $T_1(z)$ results in the CDF $H_1(z)$. Corresponding income densities $h_0(z)$ and $h_1(z)$ are shown in the bottom panel.

The horizontal distance between H_0 and H_1 quantifies the income response Δz to a reform from T_0 to T_1 at each quantile n of the distribution. Formally,

$$\Delta z(n) = H_1^{-1}(n) - H_0^{-1}(n). \tag{1}$$

This definition remains well defined even if taxpayers reorder in response to the tax reform, for example due to heterogeneous elasticities or income frictions of the kind considered in the next section. And indeed, in the presence of such heterogeneity or frictions, the response Δz is generally the statistic of primary interest for policy makers, because it quantifies the fiscal effects from tax-induced behavioral distortions. $\Delta z(n)$ can be expressed in elasticity form as[5]

$$e(n) = \frac{d\ln z}{d\ln(1-T')} = \frac{\ln\left(\frac{z+\Delta z(n)}{z}\right)}{\ln\left(\frac{1-T_1'}{1-T_0'}\right)}. \tag{3}$$

This being a static model, the CDFs H_0 and H_1 should be interpreted as steady state distributions under different counterfactual tax schedules, so that $e(n)$ represents the long-run response to a reform from T_0 to T_1.

The insight in Saez (2010) is that the observed income distribution under the piecewise-linear income tax plotted as the kinked solid line in the top panel of Figure 1,

$$T(z) := \begin{cases} T_0(z) & \text{if } z \le k, \\ T_1(z) & \text{if } z > k, \end{cases} \tag{4}$$

can provide information about the income response Δz to a reform from T_0 to T_1. This logic is illustrated in Figure 2. Suppose that each income choice is the solution to a taxpayer's frictionless optimization problem, so that their selected income represents a point of tangency between their budget constraint and an upward-sloping indifference curve arising from utility

[5]This elasticity is a weighted average of the compensated and uncompensated elasticities of taxable income $e_c(n)$ and $e_u(n)$:

$$e(n) = \left(\frac{k}{z_0(n)}\right) e_c(n) + \left(1 - \frac{k}{z_0(n)}\right) e_u(n). \tag{2}$$

where $z_0(n) := H_0^{-1}(n)$ denotes the nth quantile of the income distribution under T_0 and k is the income at which T_0 and T_1 intersect. This decomposition is derived formally in Kleven (2016), footnote 5, using the Slutsky equation.

function $u(c,z)$ which trades off the utility of consumption against the disutility of exerting effort to earn income. Then taxpayers who choose incomes below k under the kinked tax schedule T face the same local budget constraint as they would under the linear tax T_0, while taxpayers earning above k under T face the same local budget constraint as they would under the linear tax T_1. As a result, theory predicts that the observed income distribution under the kinked tax T will coincide with $H_0(z)$ at incomes below k and with $H_1(z)$ at incomes above k, with a discontinuous vertical jump at k of magnitude

$$B := H_1(k) - H_0(k), \tag{5}$$

corresponding to an atom of mass in the observed income density at k. Appendix A illustrates this logic in the presence of a tax notch, which also produces an atom of mass at the threshold k as well as a "hole" (density equal to zero) at incomes just above the bracket threshold which are utility-dominated by income k.

The vertical distance B between these CDFs is related to the horizontal distance Δz by the following equation:

$$H_0(k) + B = H_0(k + \Delta z). \tag{6}$$

Strictly speaking, this condition identifies the income response at a specific quantile of the income distribution—the quantile that earns k under the linear tax T_1, although applications of the bunching estimator approach often assume that the elasticity is constant across incomes, either globally or in a region around the kink k.

Employing equation (6), we can use a Taylor expansion of $H_0(z)$ around $z = k$ to write B in terms of Δz and the local density $h_0(k)$ and its derivatives:

$$B = h_0(k)\Delta z + \frac{h_0'(k)}{2}\Delta z^2 + \frac{h_0''(k)}{3!}\Delta z^3 + \ldots \tag{7}$$

If we regard terms of order exceeding $h_0'(k)$ as negligible (i.e., if we assume the density $h_0(z)$ is locally linear near k), then we can write the income response Δz as an explicit function of B,

$$\Delta z = \frac{B}{h_0(k) + \frac{h_0'(k)}{2}\Delta z} = \frac{B}{\left(\frac{h_0(k)+h_0(k+\Delta z)}{2}\right)}, \tag{8}$$

where the second equality uses the Taylor approximation $h_0(k+\Delta z) \approx h_0(k) + h_0'(k)\Delta z$ for small Δz. Substituting equation (8) into equation (3) yields the bunching estimator derived in Saez

(2010).[6]

This frictionless model predicts an atom of excess bunching mass in the observed income density (Figure 1, bottom panel). In contrast, such excess mass—when it is observed—is typically diffuse, like the green density line $h(z)$ in Figure 1 (see, for example, (Saez, 2010)). Equivalently, observed income CDFs generally do not jump discontinuously at tax kinks; instead, they transition gradually from $H_0(z)$ to $H_1(z)$ across a range of incomes around z, as shown by the green CDF $H(z)$. Such diffuse bunching is usually interpreted as evidence of income frictions.

In the presence of frictions, the vertical distance B between the latent CDFs $H_0(k)$ and $H_1(k)$ still quantifies the size of the key empirical parameter of interest—the steady-state income response to a marginal tax rate increase, Δz—according to equation (7). But in contrast to the frictionless setting, B cannot be estimated by measuring a vertical discontinuity in the income CDF at k.

To estimate B in the presence of frictions, Saez (2010) proposes measuring B by quantifying the excess mass in the observed density $h(z)$ around k. To formalize this logic, note that the vertical distance B is identical to the integral of the excess mass in the observed density around k, $h(z)$, relative to the counterfactual densities $h_0(z)$ and $h_1(z)$:[7]

$$B \equiv \int_{-\infty}^{k} (h(z) - h_0(z))\, dz + \int_{k}^{\infty} (h(z) - h_1(z))\, dz. \tag{9}$$

This relationship is illustrated in Figure 1, where the shaded region of excess mass in the bottom panel is equal to B.

Although this equation provides a strategy for estimating B in the presence of frictions, it requires knowledge of the counterfactual densities h_0 and h_1, one or both of which is typically unknown. Thus much of the empirical bunching literature proposes strategies for estimating excess bunching mass B from the observed density $h(z)$ without full knowledge of the counterfactual densities h_0 and h_1. For example, Chetty et al. (2011) proposes estimating B by integrating over the observed density $h(z)$ relative to an estimated counterfactual density obtained by fitting a flexible polynomial to $h(z)$ outside of a visually specified "bunching window." Kleven and Waseem (2013) extends this approach to handle notches, wherein the tax *level* (and not just the marginal tax rate) jumps discontinuously at a threshold k. While these approaches implicitly accommodate frictions by allowing for bunching to be diffuse, they leave

[6] To produce equation (5) in Saez (2010), rearrange equation (8) above to be $\Delta z = k\left[\left(\frac{1-T_1'}{1-T_0'}\right)^e - 1\right]$ and substitute it into equation (3), noting that $h_1(k) = h_0(k+\Delta z)\left(\frac{1-T_0'}{1-T_1'}\right)^e$.

[7] This identity follows from the fact that $B = H_1(k) - H_0(k) = [H(k) - H_0(k)] + [H_1(k) - H(k)]$, where the bracketed terms are equal to the two integrals in equation (9), respectively.

the process that produces diffuse bunching unmodeled.

As we show in the next section, a more explicit treatment of the diffusion process can improve the estimation of the bunching mass B in the presence of frictions by better distinguishing between the diffuse bunching that is "expected" around a kink and the underlying counterfactual densities. Moreover, by explicitly modeling these frictions and then fitting the observed shape of the diffuse mass around a kink, we can exploit additional information about income frictions that is discarded when the integrated bunching mass is reduced to a scalar measure of excess mass.

2.2 A model of sparsity-based frictions

We introduce a simple modification to the static frictionless model of Saez (2010): rather than selecting incomes from a continuum, taxpayers choose their preferred income from a sparse set of opportunities.

This notion of "sparsity-based frictions" can accommodate a diverse set of microfoundations. In a setting where incomes are produced by performing discrete jobs or gigs that are discovered via a search process, the income opportunity set represents the incomes available from the set of jobs (or combinations of jobs) that a taxpayer faces after searching with a given intensity. If an employee works for a single employer, they may face discrete choices ("lumpiness") over work shifts or overtime opportunities, rather than being able to adjust their labor hours continuously. The model can also be interpreted to allow for rational inattention, where a taxpayer learns about the precise income that arises from each potential combination of actions—such as claiming a specific set of tax deductions—only by paying information-gathering costs.[8] Taxpayers' action spaces could span both real responses—e.g., deciding which income-earning opportunities to pursue—or reporting responses—e.g., a business owner deciding which of their (lumpy) payments to realize in the current tax year, or which potentially tax-deductible expenses to claim.[9]

Formally, we assume that each taxpayer faces an *income opportunity set*,

$$\left\{z^* + \varepsilon_1,\, z^* + \varepsilon_2,\, \ldots,\, z^* + \varepsilon_M\right\}, \tag{10}$$

consisting of M income opportunities, which are offset from the taxpayer's preferred income

[8] This interpretation is in line with Jung et al. (2019), who microfound the compression of an underlying continuous distribution of actions into a lower-dimensional discrete set when information processing is costly. Such information-gathering costs have also been used in the literatures on firm price-setting and household trading in financial markets (Alvarez, Lippi and Paciello, 2011; Alvarez, Guiso and Lippi, 2012; Abel, Eberly and Panageas, 2013).

[9] For the case of lumpy tax deductions, see Rees-Jones (2018) and also discussions in the accounting literature, e.g., Kothari, Leone and Wasley (2005).

z^* by a random error term ε_i. We assume that the error terms are independent and identically distributed (iid) within taxpayer type, with distribution $F_\varepsilon(x|n)$ and density $f_\varepsilon(x|n)$. We call the combination of the error distribution F_ε and the number of draws M the *income opportunity process*.

Taxpayers of an identical "type" (productivity, elasticity, etc.) will nevertheless choose different incomes due to their different opportunities sets. The *type-conditional density* of incomes among taxpayers of a given type is characterized by the following proposition.

Proposition 1. *The type-conditional density of incomes at $\tilde{z}$ among taxpayers of type n with utility $u(c,z|n)$ under income tax $T(z)$ is given by*

$$g(\tilde{z}|n) = M \cdot f_\varepsilon\left(\tilde{z} - z^*(n)|n\right) \times \left[1 - \int_{z \in \Theta(\tilde{z}|n)} f_\varepsilon\left(z - z^*(n)|n\right) dz\right]^{M-1}, \tag{11}$$

where

$$\Theta(\tilde{z}|n) := \left\{ z \middle| u(z - T(z), z|n) \geq u(\tilde{z} - T(\tilde{z}), \tilde{z}|n) \right\}, \tag{12}$$

denoting the set of incomes that utility-dominate $\tilde{z}$ for a taxpayer of type n.

The logic behind this result is illustrated in Figure 3, which plots the indirect utility function $v(z|a)$ over income for a taxpayer of type a (see Figure 2) who faces a locally linear income tax. The type-conditional income density at income $\tilde{z}$ is equal to the probability that $\tilde{z}$ is drawn in the taxpayer's opportunity set multiplied by the probability none of the taxpayer's $M-1$ other opportunities fall in the pink region of incomes that provide higher utility than $\tilde{z}$.

The first term in equation (11), $M \cdot f_\varepsilon(\tilde{z} - z^*(n)|n)$, is the probability that $\tilde{z}$ is in an a-type's opportunity set—it is the probability of drawing the error value $\varepsilon = \tilde{z} - z^*(n)$ that would produce income $\tilde{z}$ multiplied by the M chances to draw that error value. The second (bracketed) term in equation (11) represents the probability that $\tilde{z}$ is *chosen* from the opportunity set, conditional on having drawn it. This is the probability that none of the other $M-1$ income opportunities are in the region of dominating incomes, $\Theta(\tilde{z}|n)$, shaded in pink in Figure 3. Formally, when the indirect utility function is convex, as is the case for convex preferences under a linear income tax, then $\Theta(\tilde{z}|n)$ is the interval $\left[\underline{Z}(\tilde{z}|a), \overline{Z}(\tilde{z}|a)\right]$, where $\underline{Z}(\tilde{z}|a)$ and $\overline{Z}(\tilde{z}|a)$ are the minimal and maximal values of z satisfying the equation $v(z|a) = v(\tilde{z}|a)$. The type-conditional density is maximized at the taxpayer's target income $z^*(a)$, at which point the set of dominating incomes $\Theta(z|n)$ is a singleton containing only $z^*(a)$.

Figure 4 illustrates how to extend this logic to a tax kink. Figures 4a and 4b plot the budget constraints and indifference curves for types b and c (see Figure 2) under the kinked tax schedule $T(z)$. Figures 4c and 4d plot the resulting indirect utility functions for each type, which are constructed by retaining the relevant segments of the indirect utility functions that

would obtain under each of the linear tax schedules $T_0(z)$ and $T_1(z)$.

Figure 5 illustrates the implications for the type-conditional income densities $g(z|b)$ and $g(z|c)$. The formula for the type-conditional density in Proposition 1 carries through: the effect of the tax kink operates through its effect on the set of dominating incomes $\Theta(z|n)$, which are again shaded as pink regions in Figure 5. Because the kink point k maximizes indirect utility for both types b and c (and all types in between), the type-conditional income density is also maximized at k for those types. Appendix A extends this logic to notches, where frictions lead to diffuse excess mass in the type-conditional density below the notch threshold and a density above the threshold that, while discontinuously lower, is still positive. This illustrates an important contrast between the predictions of this model, in which some taxpayers choose an income opportunity just above the notch—because all of their competing opportunities are more distant and yield lower utility—and the frictionless model, in which the density is zero above the notch, because it is strictly dominated by the threshold income k.

By aggregating these type-conditional income densities across the continuum of types $F_n(n)$, we obtain the observed income density:

$$h(z) = \int_n g(z|n) f(n) dn. \tag{13}$$

If the type density is smooth and the tax schedule is linear, then the type-conditional density effectively acts as a smoothing filter through which the type density is passed, produce an observed density that is also smooth. But because of the diffuse nature of these type-conditional densities, excess mass in the resulting income distribution is not concentrated at the kink; rather it is spread out as illustrated in the left panels of Figure 6. The right panels illustrate the effects of a notch, which produces diffuse excess mass below the notch threshold and a depression in the income density above the threshold.

2.3 A tractable case: uniform sparsity

For a given income opportunity process—combined with usual details of a frictionless bunching model—Proposition 1 allows us to compute numerically the observed income distribution. In general, this computation will depend on several structural parameters—such as the number of income opportunity draws and the parameters of the error distribution F_ε—and it may be computationally demanding.

In this section, we consider a parsimonious case, *uniform sparsity*, which is characterized by only one parameter and turns out to be a good approximation for a broad range of other income opportunity processes.

To build intuition, consider the income process simulated in Saez (1999)—the working

paper that preceded Saez (2010)—in which taxpayers have isoelastic utility

$$u(c,z|n) = c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} \tag{14}$$

and each taxpayer draws M income opportunities from a uniform distribution of width W centered around their target income $z^*(n)$.[10] This is a model of sparsity-based frictions with two additional parameters relative to the frictionless model: the number of income opportunities M and the width of the uniform distribution W from which they are drawn.

Figure 7 displays several simulated income densities that arise from this model and variations on it, plotted in the vicinity of a tax kink.[11] Panel 7a displays four simulated income densities in which each taxpayer draws $M = 1$, 2, 3, or 5 income opportunities from a uniform distribution of width $W = 50{,}000$ around their target income. When $M = 1$, the bunching mass is a rectangular plateau centered around k, which is produced by the mass of bunchers who target the income k but then draw an income opportunity that is offset from that target by a uniformly distributed error. When $M = 2$, the plateau disappears and the bunching mass approximates an inverted "V," reflecting that when taxpayers targeting income k face two opportunities, they choose the one that is closer to their target. As the number of income opportunities increases, this pattern becomes more pronounced, with a higher peak at k.

The limit of the series of densities in Figure 7a as $M \to \infty$ is the frictionless model with no diffusion in the bunching mass. However, there is an alternative notion of a limiting case in which diffusion remains non-degenerate. Consider the simulation in which $M = 5$. In this case, taxpayers' income choices are effectively determined by the distribution of just two income opportunities, the lowest opportunity above their target and the highest opportunity below their target, which dominate all other (more distant) opportunities. As a result, an income opportunity process with $M = 5$ uniform draws from a window of \$50,000 produces very similar behavior to an income opportunity process with $M = 10$ draws from a window of $W = \$100{,}000$. Both specifications produce similar distributions over the two nearest-to-target opportunities, which are uniformly drawn in the vicinity of the target with the same density $M/W = 5/50{,}000 = 10/100{,}000 = 0.0001$.

[10] Formally, the error distribution for this income opportunity process is

$$F_\varepsilon(x|n) = \begin{cases} 0 & \text{if } x < -W/2, \\ \frac{x - W/2}{W} & \text{if } -W/2 \le x \le W/2, \\ 1 & \text{if } x > W/2. \end{cases}$$

[11] These simulations use tax parameters with similar nominal magnitudes to our empirical setting: the marginal tax rate rises from $t_0 = 0.1$ to $t_1 = 0.2$ at the bracket threshold of $k = \$300{,}000$, and we assume a locally linear density of productivity n and an elasticity of taxable income of $e = 0.3$; see Section 3 for further simulation details.

Motivated by this observation, we consider the behavior of the series of income densities that arises when each agent draws M opportunities from a window of width $M \times 10{,}000$ around their target. Figure 7b plots this series with the same values of M as in Figure 7a. The income density with $M = 1$ exhibits a rectangular plateau centered around the kink, though this time with a width of \$10,000. But as M increases—and the width W increases proportionally—the bunching density appears to converge toward a distinctive "tent shape" with a peak at k.

The apparent convergence exhibited in Figure 7b motivates a natural question: does this series converge to well-defined limiting density? The answer turns out to be yes, and it is this limiting density that we call the "uniform sparsity model," formally defined as follows.

Definition 1 (Uniform sparsity)**.** *Under* ***uniform sparsity****, the income opportunity set for each taxpayer is a Poisson process with arrival rate* λ*.*

Intuitively, each taxpayer draws an infinite set of income opportunities spanning the number line. The probability of drawing any particular income opportunity is the same, with an average of λ opportunities drawn from any \$1 bin. The limiting density of the series in Figure 7b is uniform sparsity with $\lambda = 0.0001$. (For a formal proof of this claim, see Proposition 3 below.)

Under uniform sparsity, the type-conditional density has a tractable form, as shown in the following proposition.

Proposition 2. *Under the uniform sparsity model, the type-conditional density of incomes at* $\tilde{z}$ *among taxpayers of type* n *with utility* $u(c, z|n)$ *is*

$$g(\tilde{z}|n) = \lambda \exp\left[-\lambda \left|\Theta(\tilde{z}|n)\right|\right] \tag{15}$$

where $|\Theta(\tilde{z}|n)|$ *denotes the measure of the set of dominating incomes* $\Theta(\tilde{z}|n)$*, defined as in Proposition 1.*

Proof. Due to the Poisson income process, the probability that any particular income is in the income opportunity set is a constant equal to λ. The type-conditional density $g(\tilde{z}|n)$ is equal to the probability that $\tilde{z}$ is in the taxpayer's income opportunity set, which is λ, multiplied by the probability that no income opportunity is drawn from the dominating income region $\Theta(\tilde{z}|n)$, which is $\exp\left[-\int_{z\in\Theta(\tilde{z}|n)} \lambda dz\right] = \exp\left[-\lambda \left|\Theta(\tilde{z}|n)\right|\right]$. Multiplying these two terms produces the result. □

The parameter λ has a natural economic interpretation: the expected distance between adjacent income opportunities is $1/\lambda$. We call this distance the "lumpiness" of the income opportunity process. Greater lumpiness implies that taxpayers face a sparser set of income

opportunities, and thus bunching around tax kinks is more diffuse. Because it is sometimes more natural to think of these frictions as parameterized by their lumpiness rather than its inverse, we define the "lumpiness parameter" $\mu := 1/\lambda$, which has a lower bound of 0—corresponding to the frictionless model—and is unbounded above.

The convergence exhibited in Figure 7b is not unique to uniformly distributed error terms. Figure 7c displays an analogous series of income densities in the case where income opportunities are drawn from a normal (rather than uniform) distribution centered around the target income. As in Figure 7b, the spread of the distribution from which income opportunities are drawn is adjusted to preserve the density of draws in the neighborhood of the target income—this time by rescaling the standard deviation of F_ε in proportion to M. [12] As in Figure 7b, the series appears to converge quickly toward a distinctive tent shape as M increases.

Our next proposition demonstrates that this series of income densities also converges to the uniform sparsity model as $M \to \infty$. More generally, it shows that the uniform sparsity model is the limit of such a series for *any* distribution of income opportunities with positive continuous density around the income target, suggesting that this single-parameter model is a parsimonious approximation for a broad class of frictions.

To formalize this statement, we begin with an arbitrary distribution of income opportunity errors $F_\varepsilon(x)$ for which $f_\varepsilon(0) > 0$. We then define a transformation that controls the "spread" of this distribution around the target income, $F_\varepsilon^M(x) := F_\varepsilon(x/M)$, so that as M increases, the density of opportunities around the target income, $M \cdot f_\varepsilon^M(0)$ remains constant.[13] We can then show the following proposition.

Proposition 3. *For any error distribution $F_\varepsilon(x)$ with positive continuous density at $x = 0$, the income density arising from a model in which each agent draws M income opportunities offset from their target by iid disturbances $\varepsilon \sim F_\varepsilon^M$ converges pointwise to the density produced by the uniform sparsity model with $\lambda = f_\varepsilon(0)$.*

The proof is presented in Appendix B.

A striking feature of Figures 7b and 7c is that these series in M converge to the uniform sparsity model quite quickly. The simulation with just two income opportunities looks similar

[12] Formally, the probability density function of this error term distribution is $f_\varepsilon(x|n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right]$, where σ is the standard deviation of the normal distribution from which errors are drawn. The probability of a specific income opportunity draw being equal to $\tilde{z}$ is $f_\varepsilon(x|n)$, and thus the probability of *any* of the M draws being equal to $\tilde{z}$ is $M \cdot f_\varepsilon(\tilde{z}|n) = \frac{M}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right]$. Therefore the density of income opportunities in the vicinity of the target is equal to $M \cdot f_\varepsilon(0|n) = \frac{M}{\sigma\sqrt{2\pi}}$. As in the case with uniform distributions above, we jointly adjust M and the spread of the distribution—this time by adjusting the standard deviation σ—to hold constant the density of opportunities around the target income. This amounts to scaling M and σ proportionally.

[13] Note that the transformations used to construct the series of densities in Figures 7b and 7c are special cases of this general construction.

to uniform sparsity, and the simulations with $M = 3$ and $M = 5$ are nearly indiscernible.

Panels (d)–(f) of Figure 7 reproduce the simulations in Panels (a)–(c) in the case of a tax notch, where tax liability increases by \$1000 at the bracket threshold. In Panel (d)—as in Panel (a)—the case with $M = 1$ has distinctive features produced by the specifics of the uniform distribution from which income opportunities are drawn, with the bunching mass again having a plateau-like shape around the bracket threshold.[14] As the number of opportunities M increases, the mass develops a distinctive shape with diffuse mass to the left of the threshold and a depression to the right. Although convergence is slightly less rapid in the case of a notch than the case of a kink, both densities appear quite similar to the uniform sparsity model for $M = 5$.

Taken together, Proposition 3 and the simulations in Figure 7 suggest that the uniform sparsity model is a parsimonious approximation for a broad class of frictions in which taxpayers choose their final income from a sparse set containing multiple income opportunities. This model has a number of attractive features. First, the patterns of bunching it produces are strikingly similar to those observed empirically in settings with tax kinks and notches. The tent shape (high kurtosis) of the uniform sparsity specification in Figure 7b resembles both the shape of diffuse bunching observed in canonical "bunching at the kink" papers (e.g., Saez, 2010; Chetty et al., 2011; Mortenson and Whitten, 2020). Similarly, the bunching pattern around a notch in Figure 7c matches key empirical features observed in Kleven (2016), with diffuse mass to the left of the notch and a positive (but depressed) density in the "dominated region" to the right of the notch.[15] Both types of patterns appear in our empirical setting, which we discuss in Section 4.

Figure 7 also provides some reassurance that although the uniform sparsity model is not entirely general—in particular, it does a poor job of approximating the income distributions produced when $M = 1$—those cases are not the ones that appear to exhibit the key patterns of empirical bunching in many settings of interest. This is particularly evident in the case of notches, where specifications with $M = 1$ predict an income density that is *continuous* across

[14] In Panel (d), for $M = 1$, the downward slope at the left end of the plateau comes from the interaction between the mass of bunching taxpayers who target their income at the bracket threshold k and the absence of taxpayers targeting income in the dominated region to the right of k. When $M = 1$, the observed density at any particular income is simply the average of the density of target incomes—which resembles Figure 7d—in a \$50,000 window centered at that point. At \$275,000 (the left end of the plateau), this averaging window spans from \$250,000 to \$300,000, across which the target income density is positive. As income increases from \$275,000, the average falls due to the absence of target incomes to the right of k. The plateau levels out again at \$300,000 when the upper end of the averaging window rises above the upper bound of the dominated income region.

[15] Kleven and Waseem (2013) propose an alternative model that predicts positive mass in the dominated income range, in which a subset of taxpayers are insensitive to the presence of the notch due to adjustment or informational frictions. Such a model explains positive density in the dominated region, but in contrast to the uniform sparsity model, it does not predict leftward diffusion in the excess density at the bracket threshold.

the notch threshold at k, in contrast to specifications with $M > 1$, which exhibit a sharp drop in density at the notch due to the taxpayers' endogenous selection of their preferred draw. Empirical densities like those in Kleven (2016) clearly exhibit such a discontinuity, suggesting they are better represented by a model with $M > 1$, for which the uniform sparsity model performs well.[16]

A second strength of the uniform sparsity model is its parsimony. It distills the many details underlying particular sparsity-based models of frictions—such as the parametric distribution of income opportunities, the number of income opportunity draws, and the spread of the distribution (controlled, e.g., by the width of the uniform distribution or variance of the normal distribution)—into a single lumpiness parameter with a clear economic interpretation.

A third feature of the uniform sparsity model is that it has no "center," and as a result it can be conceptually separated from the taxpayer's choice of target income. This is particularly attractive in the case of notches, where a taxpayer may be indifferent between two different incomes in the frictionless model. This case is illustrated in Figure A3, where type c is exactly indifferent between earning two different incomes. Under a model of targeting or directed search, this would raise the question of which of the two equally desirable incomes should be modeled as the "target" around which opportunities are drawn. Under uniform sparsity, this question is irrelevant, because the choice of target is inconsequential.[17]

The simulations above assume the isoelastic utility function in equation (14) as in Saez (2010). In general, the model can be solved numerically for any particular specification of utility. However, if the indirect utility function is approximated by its second-order Taylor expansion around the target income, the model turns out to be particularly parsimonious. We impose this approximation in the following assumption.

Assumption 1. *Each taxpayer's indirect utility over incomes under a linear tax is approximated by its second-order Taylor approximation around their target income:*

$$v(z|n) \approx v(z^*(n)|n) + v'(z^*(n)|n)(z - z^*(n)) + \frac{1}{2}v''(z^*(n)|n)(z - z^*(n))^2. \tag{16}$$

[16] Allen et al. (2017) find bunching-based evidence of reference dependence around round numbers in the times of marathon runners, consistent with a psychological payoff for recording a time under 4 hours, for example. Notably, the bunching patterns in that paper resemble the shape of the $M = 1$ simulation in Figure 7f, suggesting that marathon times are better modeled by an imperfect targeting model with a single draw—representing one's finish time—rather than a uniform sparsity model with multiple options from which to choose.

[17] This feature also allows us to sidestep the question of whether taxpayers *anticipate* frictions when selecting their choice of target. In the case illustrated by Figure A2, for example, if type c is sophisticated, they would do better to target an income slightly below the threshold k, rather than k, in order to reduce the probability of drawing income opportunities on the "wrong side" of the notch; a naive taxpayer might instead target k, as in the Saez (1999) uncertainty model. Since the uniform sparsity model does not require specifying a target, this issue becomes irrelevant.

Assumption 1 is a useful simplification because it implies that under a linear income tax, the distance between a given income z and the taxpayer's optimal income $z^*(n)$ is the same as the distance between $z^*(n)$ and the utility-equivalent income on the "far side" of the optimum. Appendix Figure A4 demonstrates that this approximation also produces results very similar to the exact solution for the isoelastic utility function used in the simulations above. In the context of Figure 3, this assumption implies that the dominating income region shaded in pink is centered around the taxpayer's preferred income $z^*(n)$. As a result, the *measure* of the dominating income set $\Theta(\tilde{z}|n)$ under a linear tax takes a particularly simple form:

$$\Theta(\tilde{z}|n) = 2\left|z - z^*(n)\right|. \tag{17}$$

An implication of this result is that under Assumption 1 the dominating region Θ does not depend on the taxpayer's elasticity of taxable income, which governs the curvature of the indirect utility function, because under quadratic indirect utility the incomes $z^*(n) - \delta$ and $z^*(n) + \delta$ both generate the same utility for every δ, regardless of curvature. By Proposition 2, the type-conditional density $g(z|n)$ depends on a taxpayer's type n only through the dominating income set $\Theta(z|n)$, which in turn depends on type only through target income $z^*(n)$, so we can refine our notation to write the type-conditional density as $g(z|z^*)$, where z^* is the taxpayer's preferred target income under a given linear tax schedule. We will use this notation in the remainder of the paper.

The following proposition characterizes the type-conditional density of incomes under the uniform sparsity model and Assumption 1.

Proposition 4. *Under Assumption 1, under the uniform sparsity model the type-conditional density of incomes at z among taxpayers of type n under any linear income tax is given by*

$$g\left(z|z^*(n)\right) = \lambda \exp\Big[-\lambda \cdot 2\left|z - z^*(n)\right|\Big], \tag{18}$$

where $z^*(n)$ is the taxpayer's preferred income.

In words, because income opportunities are drawn uniformly from the number line, the probability that no opportunity falls in the dominating region $\Theta(z|n)$ depends only on the size of the interval, which under a linear tax is $2|z - z^*(n)|$. This probability declines exponentially at rate λ with the size of this interval, producing a tent-shaped density function centered at the taxpayer's preferred income declining exponentially in both directions. This distribution is also known as the Laplace distribution with location parameter $z^*(n)$ and scale parameter $\mu(n)$. This tractable analytic representation of the type-conditional density function significantly aids the computational estimation described in Section 2.5 below.

2.4 Identification of Model Parameters

In this Section we discuss the identification of the elasticity e and lumpiness parameter μ in our model. We follow the approach suggested in Andrews, Gentzkow and Shapiro (2020), separating our discussion in to 1) showing simulations of how different parameter values generate different income densities and 2) providing an analytical proof showing that the parameters of the uniform sparsity model are separately identified in the sense of Matzkin (2013): for a given underlying distribution of types (equivalently, target incomes), no two combinations of elasticity e and lumpiness μ will lead to the same observed income distribution.

Figure 8 provides intuition for identification by plotting simulated income densities under various combinations of values for the elasticity parameter e and the lumpiness parameter μ in the presence of a kink and a notch. Panel (a) plots simulated income densities under the baseline parameter values, as well as with lower and higher values of the elasticity e. A higher elasticity raises the overall amount of diffuse bunching mass around the kink. Panel (b) holds fixed the elasticity but varies the lumpiness parameter μ, altering the *spread* of the bunching mass around the kink while preserving the total *amount* of excess mass. Although on first glance the bunching masses in panels (a) and (b) might appear similar, on inspection the source of identification is clear. A higher elasticity e and a lower μ both lead to a higher peak at the kink, but the former achieves that higher peak by increasing the total amount of bunching mass, while the latter holds fixed the bunching mass, and thus has a much narrower spread of excess mass around the kink. Panels (c) and (d) plot such simulations in the presence of a notch.

In Appendix C we provide additional simulation results demonstrating the separate identification of e and μ, and we present a proof of conditions under which separate identification can be shown analytically. The proof proceeds in three steps. First, we demonstrate that if $h_0(z)$—the counterfactual income density that would be observed under the linear tax $T_0(z)$—is locally linear, then the corresponding density of *target incomes*, denoted $h_0^*(z)$, is given by the same locally linear function.[18] Second, we show that in a region of constant elasticity, the leftward shift in the CDF of target incomes $H_1^*(z)$ is a monotonic function of the elasticity e, and at any income above the kink, $k+\varepsilon$, the observed CDF $H(z)$ converges arbitrarily closely to the CDF $H_1(z)$ when the tax change at the kink is sufficiently small. At this point, the integrated difference between $h_0(z)$ and the observed density $h(z)$ up to $k+\varepsilon$ identifies the bunching mass B arbitrarily well, and thus the elasticity e. Finally,

[18] More generally, in Appendix Lemma 4 we show that if the observed density around income z under a linear tax is equal to a polynomial function $h(\tilde{z}) = \alpha_0 + \alpha_1(\tilde{z}-z) + \alpha_2(\tilde{z}-z)^2 + \ldots$, then the density of target incomes around z is also equal to a polynomial with coefficients $\alpha_n^* = \alpha_n - \frac{\alpha_{n+2}}{(2\lambda)^2}$. By implication, if the observed density is linear (i.e., $\alpha_n = 0$ for $n > 1$), then the target income density is also linear.

we show that elasticity e and the observed density at the kink $h(k)$ identify the lumpiness parameter $\mu = 1/\lambda$.

2.5 Estimation

We now describe how the parameters of this model can be estimated from empirical data. The empirical strategy is to select the model parameters that maximize the likelihood of observing a given empirical density. To do so, we search over the parameter values for the elasticity e and the lumpiness parameter μ. If desired, we can also allow the tax notch size dT to be an estimated parameter, treating it as a revealed feature of taxpayer behavior.

In order to estimate the model, we must impose some parametric structure on the ability density $f(n)$. As much of the bunching literature, (see Chetty et al., 2011) the key identifying assumption is that the ability distribution (and thus the counterfactual income density $h_0(z)$) is, in a sense to be made precise, well behaved in the vicinity of the bracket threshold k. Intuitively, this amounts to assuming that the bracket threshold is not located at a point in the income distribution that happens to coincide with a distortion in the underlying ability distribution.[19]

We operationalize this identification strategy by assuming that the ability density follows a polynomial of order Q, i.e.,

$$f(n|\theta) = \theta_0 + \theta_1 n + \theta_2 n^2 + \ldots = \sum_{q=0}^{Q} \theta_q n^q \tag{19}$$

for a vector $\theta = \{\theta_0, \theta_1, \ldots, \theta_Q\}$.

We then estimate the parameters of the model—e, μ, θ, and (if desired) dT—using maximum likelihood. Letting i index the observations in the data, with X_i denoting each observation's income, our starting point for the likelihood function is

$$L(e, \mu, dT, \theta) = \Pi_i h(X_i = z|e, \mu, dT, \theta). \tag{20}$$

Performing maximum likelihood estimation with this likelihood function will not result in an interior maximum, however, because we have imposed no constraint on the integral of the income density function $h(z|e, \mu, dT, \theta)$. For example, the solver can make equation (20) arbitrarily high by letting the polynomial intercept θ_0 become large. To address this, we can normalize the population density within a desired range $[z_{min}, z_{max}]$ around the bracket threshold (e.g., the income range reflected in the empirical support of the taxable income distribution). In principle, we could then perform maximum likelihood estimation

[19] Blomquist et al. (2021) explore this identification strategy and its limitations at length. See Moore (2022) for a discussion of what can be identified by bunching estimators without estimating the elasticity directly.

by computationally searching for the vector (e, μ, θ, dT) that solves the following constrained maximization problem:

$$\max_{e,\mu,\theta,dT} \sum_i \log h(X_i = z|e,\mu,dT,\theta) \quad \text{subject to} \quad \int_{z_{min}}^{z_{max}} h(z|e,\mu,dT,\theta)dz = 1. \tag{21}$$

This estimation can be implemented directly with raw microdata on incomes reported to the tax authority. In many settings, however, privacy or logistical constraints restrict the analyst to operate with a binned histogram of incomes; that is the usual data input in the bunching literature. The approach in equation (21) can be modified for use with binned data using interval censoring by letting i index bins (rather than observations) and replacing the maximand in equation (21) with $\sum_i H_i \log h(Z_i|e,\mu,\theta,dT)$, where (Z_i, H_i) denotes the income and frequency values for each bin i, and letting $h(Z_i)$ denote the probability density function from the model-predicted density at bin Z_i. We adopt this modification for our estimations in the simulations and empirical exercises that follow.

Computationally solving the constrained maximization problem in equation (21) presents a challenge. The likelihood function is

$$h(z|e,\mu,dT,\theta) = \int_{-\infty}^{\infty} g(z|n,e,\mu,dT) f(n|\theta) dn. \tag{22}$$

This is difficult because numerically integrating over a large grid of types n is time consuming, and the parameter space is very large when allowing for even a cubic polynomial, which we adopt as our baseline specification.

The problem can be converted into one that is numerically tractable by viewing the selection of the polynomial coefficients θ as an inner problem that is computed conditional on the other parameters, so that we can write the maximum likelihood problem as

$$\max_{e,\mu,dT} \sum_i H_i \log h(Z_i = z|e,\mu,\theta(e,\mu,dT)), \tag{23}$$

with the integration constraint (21) enforced by appropriate selection of the function $\theta(e,\mu,dT)$. If the inner function $\theta(e,\mu,dT)$ were selected to solve the constrained maximization in equation (21), then this approach would amount to concentrating out the parameter vector θ. For numerical expediency, we instead exploit the structure of the problem in a way that allows us to compute $\theta(e,\mu,dT)$ very quickly using polynomial regression. In effect, we select θ to minimize the sum of squared differences between the observed histogram (normalized to

sum to one) and the predicted income density:

$$\theta(e,\mu,dT) = \min_{\theta} \sum_i \left(\frac{H_i}{\sum_j H_j} - h(Z_i|e,\mu,dT) \right)^2. \tag{24}$$

To illustrate, this problem can be written in regression form as follows for the case in which $f(n|\theta)$ is cubic, where the θ coefficients are selected to minimize the sum of squared residuals $\sum_i \varepsilon_i^2$:

$$\begin{aligned}
\frac{H_i}{\sum_j H_j} &= h(Z_i|e,\mu,dT) + \varepsilon_i \\
&= \int_{-\infty}^{\infty} g(Z_i|n|e,\mu,dT) f(n|\theta) dn + \varepsilon_i \\
&= \int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) \left(\theta_0 + \theta_1 n + \theta_2 n^2 + \theta_3 n^3\right) dn + \varepsilon_i \\
&= \left[\int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) dn\right] \theta_0 + \left[\int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) n dn\right] \theta_1 \\
&\quad + \left[\int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) n^2 dn\right] \theta_2 + \left[\int_{-\infty}^{\infty} g(Z_i|n,e,\mu,dT) n^3 dn\right] \theta_3 + \varepsilon_i.
\end{aligned} \tag{25}$$

The terms in brackets require only a single numerical computation of $g(z|n,e,\mu,dT)$, after which the θ polynomial coefficients can be calculated efficiently using standard matrix inversion. This facilitates rapidly computing equation (23) searching over only the three parameters e, μ, and dT. The integration constraint in equation (21) can be enforced by using a two-step procedure in which, after selecting a provisional θ vector to solve equation (24), we adjust the intercept θ_0 so that the constraint holds exactly.

In spirit, this method resembles the widely used approach—proposed in Chetty et al. (2011)—of fitting a flexible polynomial to the observed income distribution outside of a selected "bunching window," although two differences should be noted. First, by structurally accounting for the distortion pattern produced by the bracket threshold, we need not select (visually or otherwise) an excluded "bunching window" around the threshold when computing the best-fit values of θ. Instead, even data near the bracket threshold helps identify θ. This logic suggests that this estimation method may be more robust to choices about the polynomial degree Q than standard bunching estimators, where additional flexibility may attempt to fit excess mass that spills outside the excluded bunching window. We confirm this reasoning in simulations below.

Second, this approach assumes that the smooth polynomial structure is a feature of the underlying ability distribution, $f(n)$, rather than of the observed income distribution outside the bunching window. As illustrated by Figure 1, the frictionless model actually predicts a

discontinuity in the income density around the bracket threshold due to the jump in types and the condensed mapping from types to income under higher marginal tax rates. By estimating the polynomial coefficients on the type distribution directly, this approach does not impose smoothness across that threshold.

Having implemented this maximum likelihood estimation, we can compute standard errors for our estimates using the standard maximum likelihood estimator. In our empirical application, we verify that this procedure produces results very similar to the standard errors produced using a bootstrapping procedure.

3 Simulations

Using simulated data with known underlying parameters, we can assess the performance of our proposed estimation method, as compared to the conventional approach, in the presence of sparsity-based frictions. Given the large literature estimating elasticities from kinks, as opposed to notches, we focus primarily on the conventional "kink-based" bunching estimators as in Saez (2010) and Chetty et al. (2011). We discuss the application of the notch-based estimation methods from Kleven and Waseem (2013) briefly here and in detail in Appendix G.

We specify a simulated tax kink using the same parameters as in Figures 7 and 8: the marginal tax rate rises from $t_0 = 0.1$ to $t_1 = 0.2$ at the threshold $k = 300{,}000$. We simulate income densities assuming a baseline elasticity of $e_0 = 0.3$ and a lumpiness parameter of $\mu_0 = 10{,}000$, where the "0" subscript denotes the true parameters of the data-generating process, as distinct from model estimates of the parameters, which are denoted $\hat{e}$ and $\hat{\mu}$. We construct these simulations using linear underlying ability density, $f(n|\theta) = \theta_0 + \theta_1 n$, with $\theta_0 = 1000$ and $\theta_1 = -50$. Each simulation uses a taxpayer population of 100,000, which produces an amount of sampling noise similar to our empirical distributions in Figure 11. (The simulations in Figures 7 and 8 used a much higher population size of 2 million to illustrate the shape of the bunching mass with less sampling noise.)

We simulate these income distributions in two steps. First, we draw ability values (n_i) from the known ability density $f(n|\theta)$ in the vicinity of the tax bracket threshold.[20] For each ability draw, we then simulate a set of income opportunities drawn from a Poisson process, from which we choose, for each agent, the highest-utility option.[21]

[20] Specifically, we draw 100,000 values of n_i between a set of bounds $\underline{n}$ and $\overline{n}$, with the probability of drawing any value n proportional to $f(n|\theta)$. To choose the lower bound $\underline{n}$, we note that due to frictions, the set of agents who earn a given z will include types whose target incomes are well below and well above z. Therefore, to simulate the income density near the bounds of an income range $[\underline{z}, \overline{z}]$, we must draw from an ability density with target incomes well outside that range. We choose $\underline{n}$ and $\overline{n}$ such that $z^*(\underline{n}) = \underline{z} - 100{,}000$ and $z^*(\overline{n}) = \overline{z} + 100{,}000$.

[21] To simulate income opportunity sets, we exploit the fact that differences between adjacent elements in a Poisson process are iid draws from an exponential distribution with mean μ. Thus, we can construct a random

3.1 Performance of our estimator and the conventional bunching estimator

To assess the performance of our estimation method, we simulate many rounds of data from the same data-generating process with sparsity-based frictions, and in each case, we apply our estimation procedure to jointly estimate $\hat{e}$ and $\hat{\mu}$. We are interested in whether the distribution of these estimates is centered around the parameters of the data-generating process e_0 and μ_0, and how often the estimated confidence intervals contain the true value.

One example round of simulated data is displayed in Figure 9a. The green dots plot the simulated income histogram. The estimated parameters $\hat{e}$ and $\hat{\mu}$ resulting from our maximum likelihood estimation are reported in the upper corner, along with the 95 percent confidence interval for each estimate. The orange line plots the model-predicted income density under these estimated parameter values.

We then apply the conventional bunching estimator to the same data. We implement the estimator as described in Chetty et al. (2011), except instead of selecting a bunching window via visual inspection, we employ the algorithmic method proposed by Bosch, Dekker and Strohmaier (2020). The details of this implementation are described in Appendix D. Figure 9b presents the results. The bunching window is bounded by dashed lines, and the orange line displays the fitted counterfactual density outside that window.[22] The estimated elasticity and bootstrap-based 95 percent confidence interval is reported in the corner.

Comparing the elasticity estimates from the two methods, we note that the conventional bunching estimator in Panel (b) underestimates the true elasticity of the data-generating process by 25 percent. It also provides a misleading sense of precision: the 95 percent confidence interval does not contain the true elasticity. In contrast, the sparsity-based friction estimator in Panel (a) is close to the true value of $e_0 = 0.3$, which is spanned by the 95 percent confidence interval.

To compare the relative performance of these estimators more generally, we apply them to 1000 different rounds of simulated data. Figure 10a plots the histogram of elasticity estimates from the conventional bunching estimator and from our proposed estimation method. Consistent with the results from the single simulation round, the distribution of elasticity estimates from the conventional bunching estimator lies substantially below the true elasticity

income opportunity set spanning an arbitrarily wide range around a type's preferred income $z^*(n)$ by joining a random set of above-target opportunities, $\{z^*(n)+\varepsilon_a, z^*(n)+\varepsilon_a+\varepsilon_b, z^*(n)+\varepsilon_a+\varepsilon_b+\varepsilon_c, \ldots\}$, with a random set of below-target opportunities $\{z^*(n)-\varepsilon_i, z^*(n)-\varepsilon_i-\varepsilon_j, z^*(n)-\varepsilon_i-\varepsilon_j-\varepsilon_k, \ldots\}$, where the ε values are iid draws from an exponential distribution with mean μ. In the context of a kink, where indirect utility functions are concave, only a single element must be drawn in each set, since more distant draws are guaranteed to yield lower utility. For a notch, with non-concave indirect utility functions, a larger number of opportunities is drawn, such that each agent's range of income opportunities spans across the local maxima in their indirect utility functions.

[22] Strictly speaking, this line represents the counterfactual *frequency*, equal to the counterfactual density scaled up by the bin width of the empirical histogram in order to render the plots visually comparable.

e_0. The average of elasticity estimates under the conventional approach is 0.243, and the bootstrap-based 95 percent confidence intervals contain the true e_0 in less than 10 percent of the cases. In contrast, the distribution of elasticity estimates from our proposed estimation method is centered around e_0, with an average value of $\hat{e}$ across these simulation rounds of 0.307. The estimated confidence intervals from our approach also provide an accurate sense of precision: across the 1000 estimation rounds, the 95 percent confidence intervals contained the true e_0 in 95.3 percent of cases.

The downward bias in the conventional bunching estimator appears to be driven by frictions, as illustrated by Figure 10b. To construct this figure, we reproduce distributions like those in Figure 9a using several different values of the lumpiness parameter μ_0. Figure 10b plots the mean and the 95 percent quantile interval of each distribution at each value of μ. When the lumpiness parameter is small—approaching the continuous-income-choice model—the mean estimate of $\hat{e}$ under the conventional approach is close to the true value of $e_0 = 0.3$. However, as μ_0 rises, the conventional estimator exhibits substantial bias, underestimating the true parameter by more than 50 percent at the highest plotted value of μ_0. These estimates also provide a misleading sense of precision: the 95 percent quantile intervals remain about the same size as μ_0 rises, and their upper bound falls far below e_0. In contrast, under our method, the distribution of $\hat{e}$ remains centered around e_0 as frictions increase. The 95 percent quantile interval grows with μ_0, reflecting the increasing imprecision in the elasticity estimate as lumpiness increases. This imprecision accurately reflects the greater difficulty of discerning diffuse bunching mass from underlying features of the smooth ability density when frictions are substantial.

Why do frictions cause the conventional bunching estimator to be biased downward? We highlight two contributing factors. The first arises because diffusion in the bunching mass makes it difficult to distinguish excess mass from patterns in the counterfactual income density. In the uniform sparsity model of frictions—and in many of the other sparsity-based friction models it approximates, such as when income opportunities are drawn from a normal distribution around the target income—there is no window outside of which the bunching mass falls to zero. As a result, some excess bunching mass will spill over outside of any particular bunching window—including the window chosen visually or algorithmically when implementing the conventional approach. This spillover mass tends to "pull up" the estimated polynomial fit in the vicinity of the kink, causing the procedure to underestimate the difference between the observed density and the counterfactual, and hence the bunching mass. In our model, in contrast, the distortions due to frictions are endogenously modeled throughout the income distribution, including at points far from the threshold, and so they should not exert an upward pull on the ability density around the threshold.

To explore the role of this factor in producing the bias evident in Figure 10, we note that this source of bias should become more severe when the polynomial fit is allowed to be more flexible. In Appendix Figure A7, we reproduce the estimates in Figure 9 with different polynomial degrees of 1 (linear), 3, 5, and 10. Consistent with this story, Figure A7b shows that when the polynomial degree is higher, the counterfactual density bends farther up into the bunching mass, and the elasticity estimate is more severely biased downward. In contrast, Figure A7a demonstrates that our proposed method continues to estimate $\hat{e}$ close to e_0 across all polynomial degrees, suggesting that this method is robust to misspecification in the shape of the ability density in a way that the conventional approach is not.

Although this first factor appears to play an important role in the downward bias of the conventional bunching estimator, it does not appear to be the sole explanation, because even the linear polynomial specification in Figure 9a produces a substantial underestimate of the true elasticity.

The second factor contributing to downward bias in the conventional method relates to the integration constraint imposed when estimating the counterfactual polynomial fit. The logic for such a constraint comes from the observation that any taxpayers bunching around a threshold must come from points to the right of the threshold under the counterfactual, and so the total population under the actual and counterfactual income densities must be the same.[23] However, as illustrated by the hollow blue points in Figure A1a, the presence of a kink may induce taxpayers to appear inside the plotted region who were previously outside of it. In other words, although such an integration constraint does apply to the global income density, it need not apply within the particular region over which the bunching estimator is applied.[24]

To explore the role of the integration constraint, Appendix F reproduces the results in Figure A7b using three alternative methods for fitting the counterfactual density. The first imposes a constant counterfactual density on each side of the threshold, as in Saez (2010). The second implements the Chetty et al. (2011) method described in Appendix D but without the integration constraint. The third fits a separate linear density on each side of the threshold, allowing for a break at the threshold itself. Methods 2 and 3, which were explored

[23] Describing the rationale for imposing the integration constraint, Chetty et al. (2011) remarks that an unadjusted polynomial fit "... overestimates [the bunching mass] because it does not account for the fact that the additional individuals at the kink come from points to the right of the kink. That is, it does not satisfy the constraint that the area under the counterfactual must equal the area under the empirical distribution. To account for this problem, we shift the counterfactual distribution to the right of the kink upward until it satisfies the integration constraint."

[24] Indeed, Figure A1b, which illustrates the observed income density in the frictionless model with a continuous uniform type density, demonstrates that the kink may induce both extra mass at the kink *and* higher density at incomes above the kink, in which case the true counterfactual density under T_0—which extends the uniform density below k to points above it—clearly has a lower integral over the plotted region than the observed density does.

in Mortenson and Whitten (2016)—the working paper that preceded Mortenson and Whitten (2020)—substantially reduce the bias in the elasticity estimate when frictions are small to medium. Indeed, the integration constraint appears to be the source of the (slight) downward bias in the conventional estimator at low values of μ. At the same time, this factor does not fully account for the downward bias in the presence of frictions, as even these specifications without the integration constraint produce severely downward-biased elasticity estimates when frictions are more pronounced at $\mu = 15$ and $\mu = 30$. They exhibit downward bias similar in magnitude to the Chetty et al. (2011) method with the integration constraint.

3.2 Performance of the notch-based bunching estimator

We can use a similar procedure to examine the behavior of the notch-based elasticity estimator from Kleven and Waseem (2013) (abbreviated KW) in the presence of sparsity-based frictions.

The KW estimator assumes a model of frictions that is different from the sparsity-based frictions considered in this paper. In their model, a subset of agents are unresponsive to the presence of the notch, explaining the presence of mass in the dominated income range above the tax bracket threshold. The prevalence of such unresponsive agents can be found by computing the ratio of the empirical density of taxpayers in the dominated income range to the estimated counterfactual density, absent a notch, in that range. In such settings, this KW "unresponsiveness share" is a non-parametric quantification of frictions. To estimate the structural elasticity, the KW method scales up the observed excess mass to compute the bunching that would arise if all taxpayers overcame their frictions.

Sparsity-based frictions provide an alternative explanation for the presence of taxpayers with incomes in the dominated range. In Appendix G, we examine the behavior of the KW notch-based estimator applied to data from simulations that assume sparsity-based frictions. In these simulations, the KW method produces elasticity estimates that are higher than the structural elasticity of the data-generating process. This overestimate is driven by the KW rescaling of the bunching mass to account for unresponsive taxpayers. This arises from the different microfoundations underlying the two approaches. In a setting with sparsity-based frictions, mass in the dominated income range is produced by the distribution of sparse income opportunities, rather than by a share of unresponsive taxpayers; scaling up the bunching mass thus overestimates the structural elasticity.

In summary, our procedure complements Kleven and Waseem (2013) by providing an additional notch-based estimator based on an alternative model of frictions. Because the models predict somewhat different patterns of bunching around a notch, the choice between

them can potentially be informed by the data.[25]

4 Empirical application

We apply our estimation method using comprehensive administrative data from the South African Revenue Service. We apply the method to estimate the elasticity among small businesses and individuals, studying their behaviour using data from 2014 to 2018. For firms, we apply the method to study the distribution of reported corporate income tax (CIT) returns around three prominent tax kinks in the Small Business Corporation tax schedule. For individuals, we focus on the reported income distribution obtained from personal income tax (PIT) returns around kinks in the PIT schedule. The piecewise-linear tax schedules for each population are described below; bunching patterns around each kink are displayed in Figures 11 and 13.

4.1 Setting: small business and individual taxation in South Africa

4.1.1 Small business taxation in South Africa

Like many developing countries, South Africa relies more heavily on corporate income taxes than most developed economies. In 2017, corporate taxes accounted for 16.2 percent of total tax revenue in South Africa, considerably higher than the OECD average of 9.7 percent, but in line with the average shares for Africa (18.5 percent) and Latin America (15.4 percent).[26] The South African corporate tax system is tiered, with a progressive, kinked tax schedule for "Small Business Corporations" (SBCs), and a flat 28 percent tax applying to other (larger) resident corporations.[27] Corporate taxable income comprises gross revenues, less non-capital expenses and incurred losses from previous tax years which can be carried forward.[28] There are no local

[25] There are three differing predictions about the bunching mass which might be used to choose between these models of frictions. The KW model predicts (1) tight bunching at the bracket threshold among the subset of responsive tax payers, (2) upward-sloping density above the notch in the dominated income range, and (3) empirical density in the dominated range that is strictly below the estimated counterfactual density. (See KW Figure II.) The sparsity-based frictions model predicts (1) leftward diffusion in bunching at the bracket threshold, (2) U-shaped density above the notch, and (3) empirical density in the dominated range that may be above the counterfactual density. (See our Figure 8d.)

[26] Data from the OECD accessed at: `https://stats.oecd.org/Index.aspx?DataSetCode=CTS_REV`.

[27] Businesses can optionally register as an SBC if they meet a set of requirements, the most pertinent being that their annual revenue must be below R20 million (about $1.4 million US). SBCs account for 38 percent of all formally registered companies, although due to their smaller size, they account for less than 20 percent of total corporate tax revenues.

[28] Corporate dividends are taxed at the shareholder level at a 15 percent rate. We describe the full set of eligibility requirements, and other details of SBCs in Appendix H.

corporate income taxes in South Africa; businesses pay income tax only at the national level.[29]

For the purposes of this paper, we focus on SBCs because they face a piecewise-linear kinked tax schedule ideal for bunching estimation. The lowest bracket threshold, at which the marginal tax rate rises from 0 to 7 percent, is located at R75,750, or about $5,260 USD in 2018.[30] This threshold moves over time with inflation. For comparison, in 2018, South African GDP per capita was roughly $7,000 US. Below this threshold, firms face no tax liability, although they are still legally required to file a tax return.

The middle and upper thresholds are at R365,000 and R550,000, respectively, and are fixed in nominal terms. At the middle threshold, the marginal tax rate rises from 7 to 21 percent; at the upper threshold, it rises to 28 percent, meaning that firms with incomes above this threshold face the same marginal tax rate as non-SBC firms. Appendix Figure A11 (panel a) plots the schedule for 2017, and Table A1 reports the full SBC tax schedule for each year from 2010 to 2018. For our analysis, we study the population of SBCs from 2014 to 2018—this is the period over which the three-kink structure illustrated in Figure A11 was in place.

Figure 11 presents the histogram of taxable income amounts around each of the three kinks, along with the model estimates described below. Pronounced bunching is visible around each kink, with diffusion suggestive of income frictions. Interestingly, the bunching pattern around the lowest kink is distinctly asymmetric, suggestive of the patterns typically associated with a tax notch. We return to this point in our discussion of the results.

4.1.2 Personal taxation in South Africa

The personal income tax schedule in South Africa, like the schedule for SBCs, is piecewise-linear and progressive. The schedule in 2017 is plotted in Appendix Figure A11 (panel b) and we describe the rules of the personal income tax system in greater detail in Appendix H. The lowest kink point in the schedule is framed as a standard deduction—all South Africans receive a tax deduction equal to the amount of taxable income at the lowest threshold (R70,000 or $4,860 per annum in 2017). This means that any income below this threshold is taxed at a 0% marginal tax rate. The brackets are adjusted each year but do not track inflation, which leads to bracket creep across time.

Figure 13 presents the histogram of personal incomes for self-employed individuals around each of the five kinks, along with the model estimates described in the next section. Bunching is pronounced at the bottom two kinks, and visible (though less pronounced) at the third kink. No bunching is discrernable at the fourth and fifth kinks. Consistent with other results in

[29] See Pieterse, Gavin and Kreuser (2018) for more on the South African corporate income tax data.

[30] Throughout the paper, we use an exchange rate of 14.4 South African rand per U.S. dollar, which was the prevailing rate at the end of 2018.

the bunching literature, bunching is more muted among wage earners, whose histograms are displayed in Appendix Figure A12. Only the first kink exhibits evidence of bunching.

In the next subsection, we discuss estimates produced by the uniform sparsity model of frictions around each of the kink points in the South African individual and small business tax schedules, and we compare the results with those obtained using the conventional bunching estimator.

4.2 Results

4.2.1 Results from small business tax returns

Panels (a)–(c) of Figure 11 presents the estimation of the uniform sparsity model using the distribution of incomes from small businesses. The estimated elasticities of taxable income are 0.27 and 0.23 at the middle and upper kinks, respectively. The elasticity estimated at the lowest kink is substantially higher—in excess of one—consistent with the larger total excess mass around that kink.

In addition to these elasticities, Panels (a)–(c) of Figure 11 also report the estimated lumpiness parameters μ and the estimated notch values dT, which we treat as a model-estimated parameter in light of the evidence that firms at the bottom kink appear to treat the bracket threshold as-if it were a notch.

Lumpiness parameters range from R5,900 to R11,300 ($410 to $785 US). These estimates capture the size of optimization frictions faced by firms at each tax kink. Between the lower and middle kinks, μ increases with income, which is consistent with distances between lumpy income opportunities increasing with total income. However, μ estimates are non-monotonic, declining from the middle to the upper kink. As we discuss below, this pattern could arise from heterogeneity across firms in their use of tax practitioners.

In Panels (d)–(f) we compare these elasticity estimates to those produced by the conventional bunching estimator. Details of the conventional estimation method are described in Appendix D. The estimated counterfactual densities in orange.[31] The elasticity estimates from the conventional estimator are roughly 30-50 percent lower than those estimated under sparsity-based frictions, and there is no overlap between the 95 percent confidence intervals produced by the two methods for any of the kinks. These results suggest that the concerns raised in Section 3 about downward bias and potential imprecision of the conventional bunching estimator in the presence of optimization frictions could be economically sizeable.

As noted above, firm behavior around the lowest kink is suggestive of a tax notch, rather

[31] These estimates conform closely with Boonzaaier et al. (2019), who use the conventional bunching estimator approach to estimate income elasticities at each of these kinks.

than a pure kink. Figure 11a shows that the estimated notch value is R340, or about $24 US, and is highly statistically significant. The model therefore strongly rejects pure kink behavior. Interestingly, although the bunching patterns around the middle and upper kinks are less strikingly reminiscent of a notch, upon close inspection they also exhibit asymmetry in their density around the bracket threshold. Consistent with this asymmetry, the model estimates also find significantly positive notch values at these thresholds. Indeed, the notch value at the middle kink is similar in magnitude to that at the lowest kink, although its smaller magnitude relative to baseline income, and the lower elasticity at this kink, result in a less pronounced asymmetry. This suggests that the behavioral tendency to treat a kink as though it were a notch is not isolated to the lowest kink, where the tax liability changes from zero to positive, but may be a more general phenomenon. As such, it suggests that the source of this "as-if" notch value is unlikely pure behavioral aversion to paying a positive tax liability. Although identifying the source of "as-if" notch behavior is beyond the scope of our analysis, such behavior would be consistent with a subset of taxpayers mistaking the discontinuity in marginal tax rates for a discontinuity in *average* tax rates, or other frictions which produce a perceived discrete cost when one's income surpasses each kink. In contrast, the estimated notch value at the upper kink is small in magnitude—equal to about $3 US—suggesting that taxpayer behavior at that threshold is not meaningfully different from that expected around a pure kink.

We next explore heterogeneity in bunching behavior across firm attributes. Specifically, we separately estimate our model for small businesses that do and do not use a registered tax practitioner to prepare their corporate income tax returns. Figure 12 shows plots like those in Figure 11, partitioning businesses on their tax practitioner usage. The raw histograms show that there is substantially more pronounced visible bunching for firms that use tax practitioners. Table 1a reports the corresponding elasticities and other parameters, together with the estimates for the aggregate population in Figure 11. The table shows that differences in bunching behavior do not appear to be driven by a consistent difference in the income elasticity between businesses that do and do not use tax practitioners. Although the estimated elasticity is higher among firms that use tax practitioners at the lowest kink, the reverse is true at the middle kink, and at the upper kink, the elasticities are not statistically distinguishable from one another. That said, there is a clear and consistent difference between the lumpiness parameters of the different groups of firms. At every one of the kinks, our method estimates that optimization frictions are smaller for firms who use tax practitioners. This is consistent with such firms fine-tuning their incomes more precisely in response to tax incentives, or paying closer attention to the set of possible actions—real economic activity or reporting behavior—which can be used to target incomes more precisely to a desired level.[32]

[32] These estimates also help to explain the non-monotonicity in μ across incomes in Figure 11. After conditioning

Tax practitioner usage also predicts a lower "as-if" notch value: firms with paid tax preparers treat a statutory kink less like a notch than firms that prepare their own tax returns. If, as hypothesized above, "as-if" notch behavior is driven by average-versus-marginal tax rate confusion, this is consistent with tax-practitioners helping firms to clear up such confusion. This also helps explain the small size of the estimated notch value at the upper kink in the aggregate sample (Figure 11c), which seems primarily driven by businesses with paid tax practitioners.

In contrast with our approach, the conventional bunching estimator does not detect this heterogeneity in behavior. Table 1b reports the income elasticities estimated by the conventional approach at each of the three kinks. Naturally, the conventional bunching estimator cannot estimate differences in the degree of income frictions or the "as-if" notch value, since these parameters are not estimated by the conventional model. However, the income elasticities estimated using the conventional approach are also indistinguishable from one another. The behavior around the middle kink illustrates this issue. Comparing Panels (b) and (e) in Figure 12 from the new optimization frictions approach, firms with tax preparers exhibit a substantially—and statistically significantly—lower income elasticity than those without tax preparers (0.26 vs. 0.50). However, Table 1b shows that the conventional bunching estimator returns statistically indistinguishable elasticity estimates (0.13 vs. 0.11). Put differently, optimization frictions in income adjustment may lead the conventional approach to misinterpret heterogeneity in lumpiness or "as-if" notch behavior in a manner that masks real differences in elasticity parameters.

We can use this heterogeneity to explore how our model estimates across subsamples relate to the estimates produced on the full population when behavior is assumed to be homogeneous. Consider a scenario where our population of interest is divided into two sub-populations, one with low lumpiness (precise bunching kinks) and another subpopulation with high lumpiness (diffuse bunching). A natural question in such a scenario is whether our approach is able to appropriately identify the aggregate elasticity parameter in the presence of such heterogeneity in diffusion/lumpiness. A fuller answer to this will entail analytical extensions beyond the scope of this paper, but we provide some insight into this question using three exercises.

First, in Appendix Table A4 reports Mean Average Percentage Errors (MAPE) between 0.20% and 0.36% between disaggregated vs. aggregated model estimates. Put differently, the average difference between the model fit estimated on the full sample and the model fit obtained by

on tax preparer usage, the apparent decline in μ from the middle kink to the upper kink shrinks considerably and confidence intervals for these estimates overlap, suggesting that differences not explained by tax practitioner usage may not be economically meaningful.

aggregating across the two subsamples is at most 0.36 per cent, i.e., the "summed subsample" and aggregate model fits are virtually identical.

Second, we compare the elasticity and lumpiness parameters estimated on the aggregate sample with a weighted average of these parameters produced on the subsamples, where the weights are the population share of each subsample. For the elasticity, the weighted subsample elasticities are close (within the 95% confidence interval) to the elasticity estimates in the aggregate sample. This finding is somewhat reassuring because it suggests that the aggregate elasticity estimate is not highly sensitive to heterogeneity in lumpiness across subsamples. For the lumpiness parameter μ, at the lower kink, we find that the "weighted average" parameter is comparable to the μ estimated on the full distribution from the "weighted average." However, at the middle and upper kinks, the values for subsample-weighted μ exceed the 95% confidence interval of the μ estimated on the full distribution. This is perhaps not surprising, given the evident heterogeneity in diffusion that is visible in these subsample distributions.

4.2.2 Results from personal income tax returns

When applying our model estimation to personal income taxes, we estimate the model separately for self-employed individuals and wage earners, the latter of which are known to be less responsive to tax kinks and are described in Appendix H and Figure A12.

Figure 13 displays the histogram and model estimates for self-employed individuals around each of the first three kinks in the personal income tax schedule, for which some bunching appears to be evident. These individuals exhibit pronounced bunching around the first two kinks. Like small businesses, individuals around the lower kink exhibits striking "notch-like" asymmetry in bunching, with an evident discontinuity at the bracket threshold and missing mass to the right. Consistent with that hypothesis, Figure A12a shows that the estimated notch value is R310, ($22 USD), which is highly statistically significant. Accounting for this "as-if" notch behavior, we estimate an elasticity at the first kink of 0.53. No discernable bunching is apparent above the third kink—consistent either with a low elasticity or a high degree of lumpiness. We perform a similar estimation for wage earners around the first kink—the only kink at which bunching is evident—which again suggests "as-if" notch behavior (Figure A12).

5 Conclusion

This paper extends the theory underlying bunching-based elasticity estimators to incorporate a positive model of frictions, and it provides new estimation methods to recover elasticities in the presence of such frictions. We consider a general class of sparsity-based frictions in

which taxpayers select their preferred option from a sparse set of opportunities, which can be microfounded using a variety of models including directed search, limited attention, and lumpy adjustment. We show that many models within this class of frictions are well approximated by a parsimonious limiting case in which opportunities are drawn from a Poisson process governed by a single "lumpiness" parameter, which quantifies the expected distance between adjacent opportunities. This model predicts key patterns observed in empirical bunching settings, such as diffuse bunching around kinks and positive mass above notches.

Simulations suggest that conventional bunching estimation techniques exhibit limitations in the presence of sparsity-based frictions, tending to underestimate the bunching mass and elasticity while producing overly precise confidence intervals. We propose an alternative estimation method that recovers unbiased elasticity estimates in simulated data. We show that this approach can be used to draw novel economic insights. When we apply this method to administrative tax data on small firms and individuals around kinks in both the corporate and personal income tax schedule in South Africa, we find evidence that both firms and individuals treat the lowest tax kink like a notch, and we estimate substantially lower income frictions among firms with paid tax preparers, consistent with finer income targeting among that group. By quantifying the extent of frictions, our approach allows us to recover the "as-if" notch value when taxpayers treat a kink like a notch.

Although we focus on the setting of earned income, the model and methods presented here are versatile. Our proposed bunching estimator can be applied to estimate behavioral responses in other settings with kinked budget sets, including with non-income tax instruments, nonlinear pricing schedules, or non-monetary payoffs. More generally, the model of uniform sparsity, as an approximation of sparsity-based frictions, can be extended to a wide range of settings, including multidimensional choices.

References

Abel, Andrew B, Janice C Eberly and Stavros Panageas. 2013. "Optimal inattention to the stock market with information costs and transactions costs." *Econometrica* 81(4):1455–1481.

Allen, Eric J, Patricia M Dechow, Devin G Pope and George Wu. 2017. "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Science* 63(6):1657–1672.

Alvarez, Fernando E, Francesco Lippi and Luigi Paciello. 2011. "Optimal price setting with observation and menu costs." *The Quarterly Journal of Economics* 126(4):1909–1960.

Alvarez, Fernando, Luigi Guiso and Francesco Lippi. 2012. "Durable consumption and asset management with transaction and observation costs." *American Economic Review* 102(5):2272–2300.

Andersen, Steffen, Cristian Badarinza, Lu Liu, Julie Marx and Tarun Ramadorai. 2022. "Reference Dependence in the Housing Market." *American Economic Review* 173(10):3398–3440.

Andrews, Isaiah, Matthew Gentzkow and Jesse M Shapiro. 2020. "Transparency in structural research." *Journal of Business & Economic Statistics* 38(4):711–722.

Bachas, Pierre and Mauricio Soto. 2021. "Corporate Taxation under Weak Enforcement." *American Economic Journal: Economic Policy* 13(4):36–71.

Best, Michael Carlos, Anne Brockmeyer, Henrik Jacobsen Kleven, Johannes Spinnewijn and Mazhar Waseem. 2015. "Production versus Revenue Efficiency with Limited Tax Capacity: Theory and Evidence from Pakistan." *Journal of Political Economy* 123(6):1311–1355.

Blomquist, Sören, Whitney K. Newey, Anil Kumar and Che-Yuan Liang. 2021. "On Bunching and Identification of the Taxable Income Elasticity." *Journal of Political Economy* 129(8):2320–2343.

Boonzaaier, Wian, Jarkko Harju, Tuomas Matikka and Jukka Pirttilä. 2019. "How Do Small Firms Respond to Tax Schedule Discontinuities? Evidence from South African Tax Registers." *International Tax and Public Finance* 26(5):1104–1136.

Bosch, Nicole, Vincent Dekker and Kristina Strohmaier. 2020. "A Data-Driven Procedure to Determine the Bunching Window: An Application to the Netherlands." *International Tax and Public Finance* 27(4):951–979.

Brehm, Margaret, Scott A Imberman and Michael F Lovenheim. 2017. "Achievement Effects of Individual Performance Incentives in a Teacher Merit Pay Tournament." *Labour Economics* 44:133–150.

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

Chetty, Raj, John N. Friedman, Tore Olsen and Luigi Pistaferri. 2011. "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *The Quarterly Journal of Economics* 126(2):749–804.

Dee, Thomas S, Will Dobbie, Brian A Jacob and Jonah Rockoff. 2019. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations." *American Economic Journal: Applied Economics* 11(3):382–423.

Devereux, Michael P., Li Liu and Simon Loretz. 2014. "The Elasticity of Corporate Taxable Income: New Evidence from UK Tax Records." *American Economic Journal: Economic Policy* 6(2):19–53.

Diamond, Rebecca and Petra Persson. 2016. "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests." *Working Paper no. 22207, National Bureau of Economic Research* .

Feldstein, Martin. 1999. "Tax Avoidance and the Deadweight Loss of the Income Tax." *Review of Economics and Statistics* 81(4):674–680.

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *The Quarterly Journal of Economics* 129(4):1661–1710.

Gelber, Alexander M., Damon Jones and Daniel W. Sacks. 2020. "Estimating Adjustment Frictions Using Nonlinear Budget Sets: Method and Evidence from the Earnings Test." *American Economic Journal: Applied Economics* 12(1):1–31.

Gordon, Roger and Wei Li. 2009. "Tax Structures in Developing Countries: Many Puzzles and a Possible Explanation." *Journal of Public Economics* 93(7-8):855–866.

Grubb, Michael D and Matthew Osborne. 2015. "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock." *American Economic Review* 105(1):234–271.

Ito, Koichiro. 2014. "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing." *American Economic Review* 104(2):537–563.

Jung, Junehyuk, Jeong Ho Kim, Filip Matějka and Christopher A Sims. 2019. "Discrete actions in information-constrained decision problems." *The Review of Economic Studies* 86(6):2643–2667.

Kemp, Johannes Hermanus. 2019. "The elasticity of taxable income: The case of South Africa." *South African Journal of Economics* 87(4):417–449.

Kleven, Henrik J. and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *The Quarterly Journal of Economics* 128(2):669–723.

Kleven, Henrik Jacobsen. 2016. "Bunching." *Annual Review of Economics* 8:435–464.

Kostøl, Andreas R. and Andreas S. Myhre. 2021. "Labor Supply Responses to Learning the Tax and Benefit Schedule." *American Economic Review* 111(11):3733–3766.

Kothari, S.P., Andrew J. Leone and Charles E. Wasley. 2005. "Performance Matched Discretionary Accrual Measures." *Journal of Accounting and Economics* 39(1):163–197.

Lediga, Collen, Nadine Riedel and Kristina Strohmaier. 2019. "The elasticity of corporate taxable income—Evidence from South Africa." *Economics Letters* 175:43–46.

Liu, Li and Ben Lockwood. 2015. VAT notches. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association.* Vol. 108 JSTOR pp. 1–51.

Manoli, Day and Andrea Weber. 2016. "Nonparametric Evidence on the Effects of Financial Incentives on Retirement Decisions." *American Economic Journal: Economic Policy* 8(4):160–182.

Matzkin, Rosa L. 2013. "Nonparametric identification in structural economic models." *Annu. Rev. Econ.* 5(1):457–486.

Mavrokonstantis, Panos and Arthur Seibold. 2022. "Bunching and Adjustment Costs: Evidence from Cypriot Tax Reforms." *Journal of Public Economics* 214:104727.

Moore, Dylan. 2022. "Evaluating Tax Reforms without Elasticities: What Bunching Can Identify." *Working paper* .

Mortenson, Jacob A. and Andrew Whitten. 2016. "How Sensitive Are Taxpayers to Marginal Tax Rates? Evidence from Income Bunching in the United States." *Working Paper* .

Mortenson, Jacob A. and Andrew Whitten. 2020. "Bunching to Maximize Tax Credits: Evidence from Kinks in the US Tax Schedule." *American Economic Journal: Economic Policy* 12(3):402–432.

Pieterse, Duncan, Elizabeth Gavin and C. Friedrich Kreuser. 2018. "Introduction to the South African Revenue Service and National Treasury Firm-Level Panel." *South African Journal of Economics* 86:6–39.

Pillay, Neryvia. 2021. "Taxpayer responsiveness to taxation: Evidence from bunching at kink points of the South African income tax schedule." *Working Paper* .

Rees-Jones, Alex. 2018. "Quantifying Loss-Averse Tax Manipulation." *The Review of Economic Studies* 85(2):1251–1278.

Rees-Jones, Alex and Dmitry Taubinsky. 2020. "Measuring "Schmeduling"." *The Review of Economic Studies* 87(5):2399–2438.

Saez, Emmanuel. 1999. "Do Taxpayers Bunch at Kink Points?" *Working Paper no. 7366, National Bureau of Economic Research* .

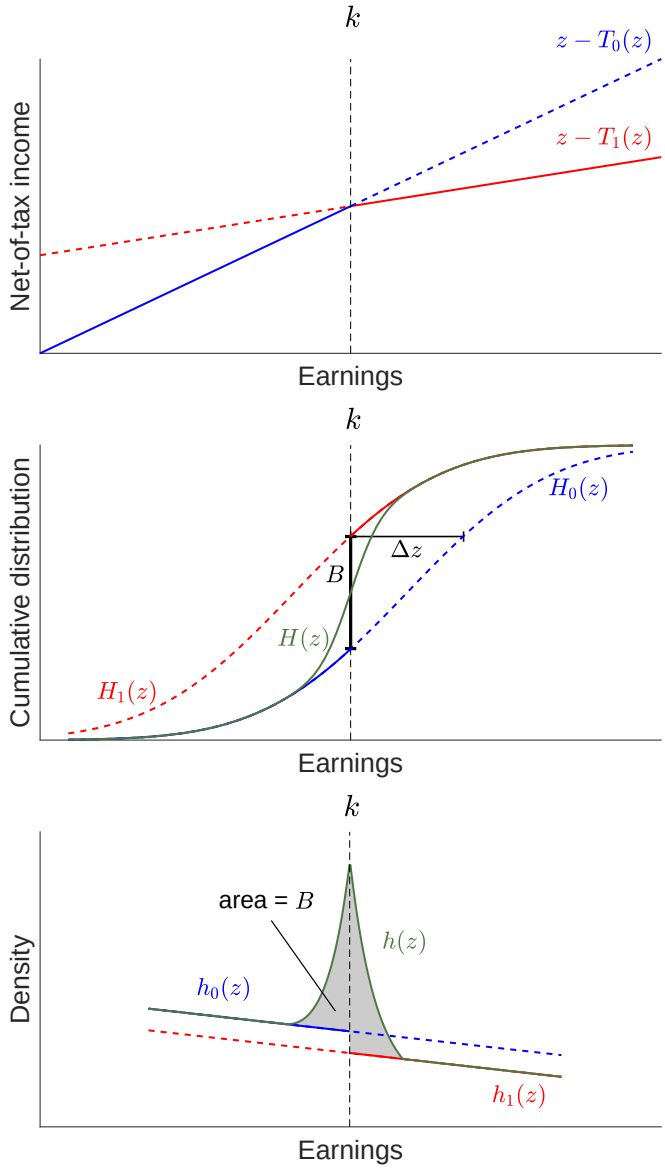
Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* 2(3):180–212.

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50(3):665–690.

Søgaard, Jakob Egholt. 2019. "Labor Supply and Optimization Frictions: Evidence from the Danish Student Labor Market." *Journal of Public Economics* 173:125–138.

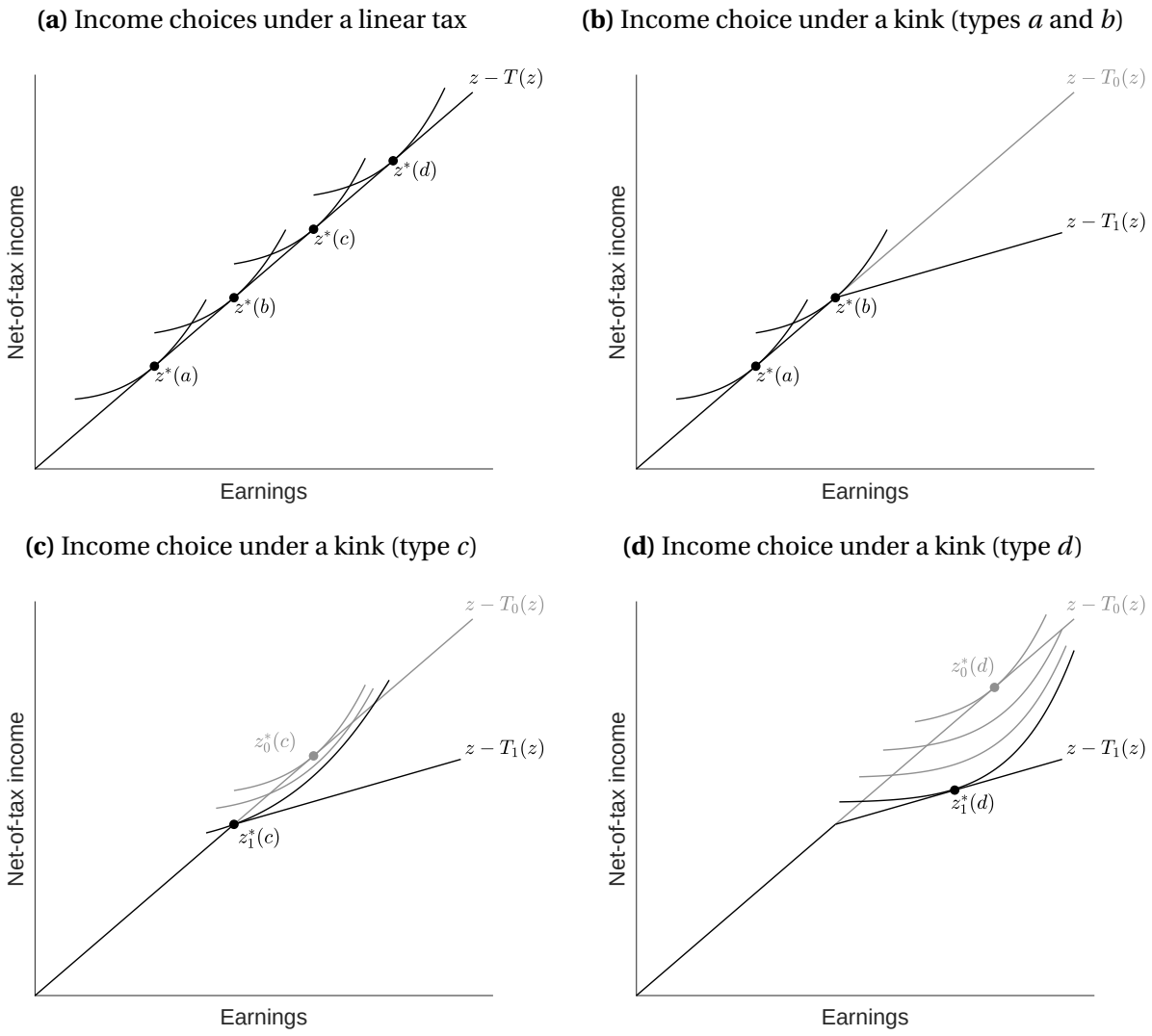
Velayudhan, Tejaswi. 2018. Misallocation or misreporting? evidence from a value added tax notch in india. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association.* Vol. 111 JSTOR pp. 1–46.

Figure 1: Income Distributions and Densities Under a Kinked Tax Schedule



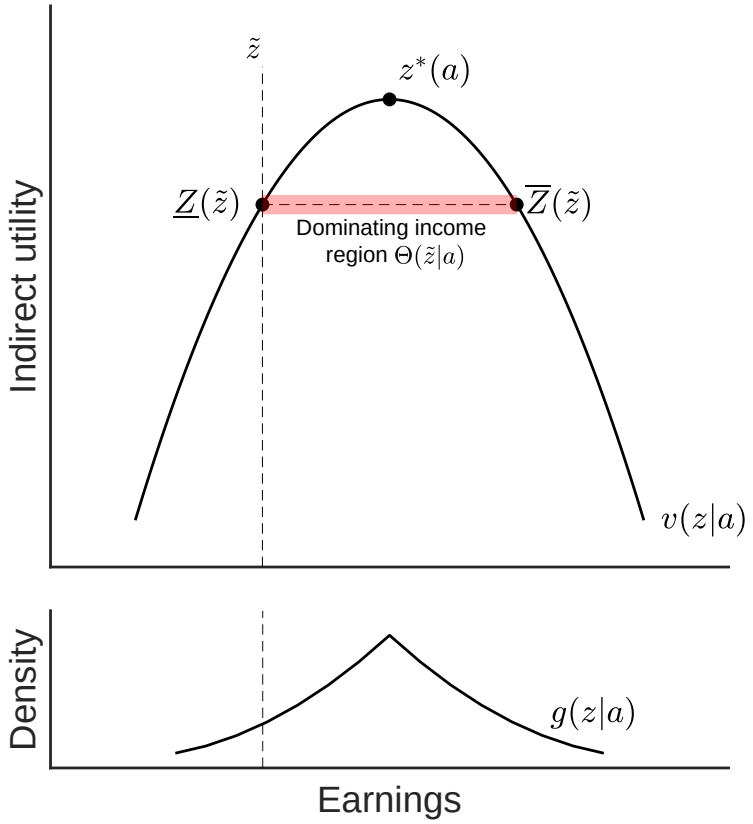
The top panel shows a taxpayer's budget constraints under two linear income tax functions, $T_0(z)$ and $T_1(z)$. These can construct a "kinked" tax schedule (with resulting budget constraint plotted by the solid line) consisting of $T_0(z)$ and $T_1(z)$ below and above income k, respectively. The middle panel plots the income CDFs $H_0(z)$ and $H_1(z)$ that would arise under each linear tax functions. Absent frictions, the CDF under the kinked tax schedule coincides with $H_0(z)$ below k and with $H_1(z)$ above k, with a discontinuous jump B at the threshold k. With frictions, the transition from $H_0(z)$ to $H_1(z)$ around the kink at k is gradual, as plotted by the CDF labeled $H(z)$. The bottom panel plots the income densities $h_0(z)$, $h_1(z)$, and $h(z)$, corresponding to the CDFs $H_0(z)$, $H_1(z)$, and $H(z)$, respectively.

Figure 2: Frictionless bunching model with a progressive tax kink



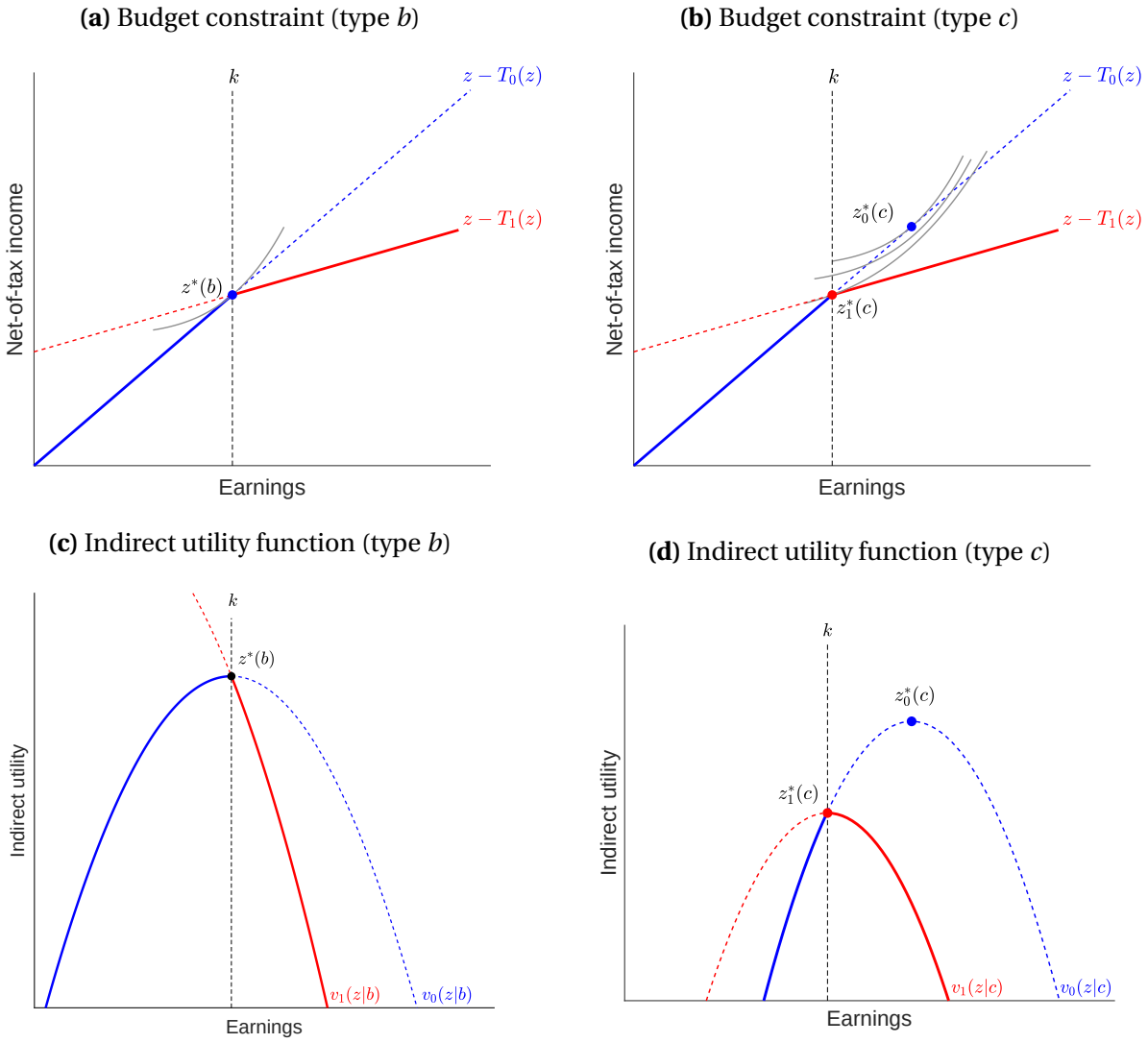
This figure illustrates the income choice around a tax kink under the conventional frictionless model. Panel (a) illustrates the optimal choice of income, z^*, for four selected types of taxpayers under a linear income tax. Panels (b), (c), and (d) illustrate the optimal choice for each type under a kinked income tax, where the tax changes from the $T_0(z)$ to $T_1(z)$ at the threshold k. Incomes z_0^* and z_1^* denote the optimal choice under the linear tax $T_0(z)$ and $T_1(z)$, respectively.

Figure 3: Type-conditional income density under a linear tax (type a)



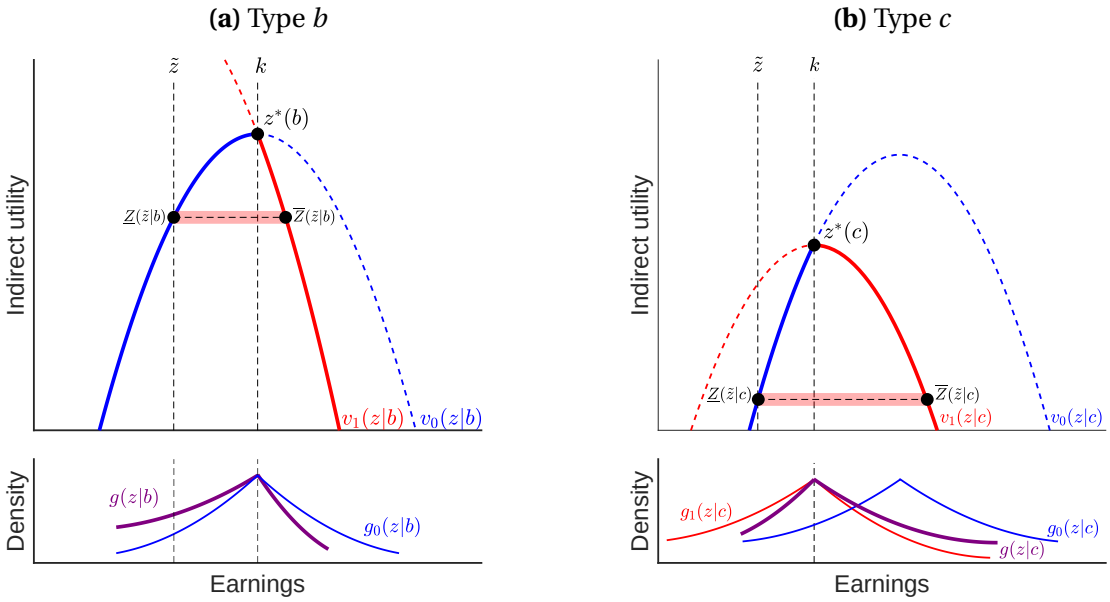
This figure illustrates the calculation of the type-conditional income density in the uniform sparsity model among a-type agents at a particular income level $\tilde{z}$, under a locally linear income tax. The top portion plots the taxpayer's indirect utility function over incomes. An agent who has $\tilde{z}$ in their income opportunity set will select this income iff they do not have some other income opportunity in the shaded "dominating region." The type-conditional density $g(\tilde{z}|a)$ is equal to this conditional probability multiplied by the probability of drawing $\tilde{z}$.

Figure 4: Utility from income choices around a tax kink



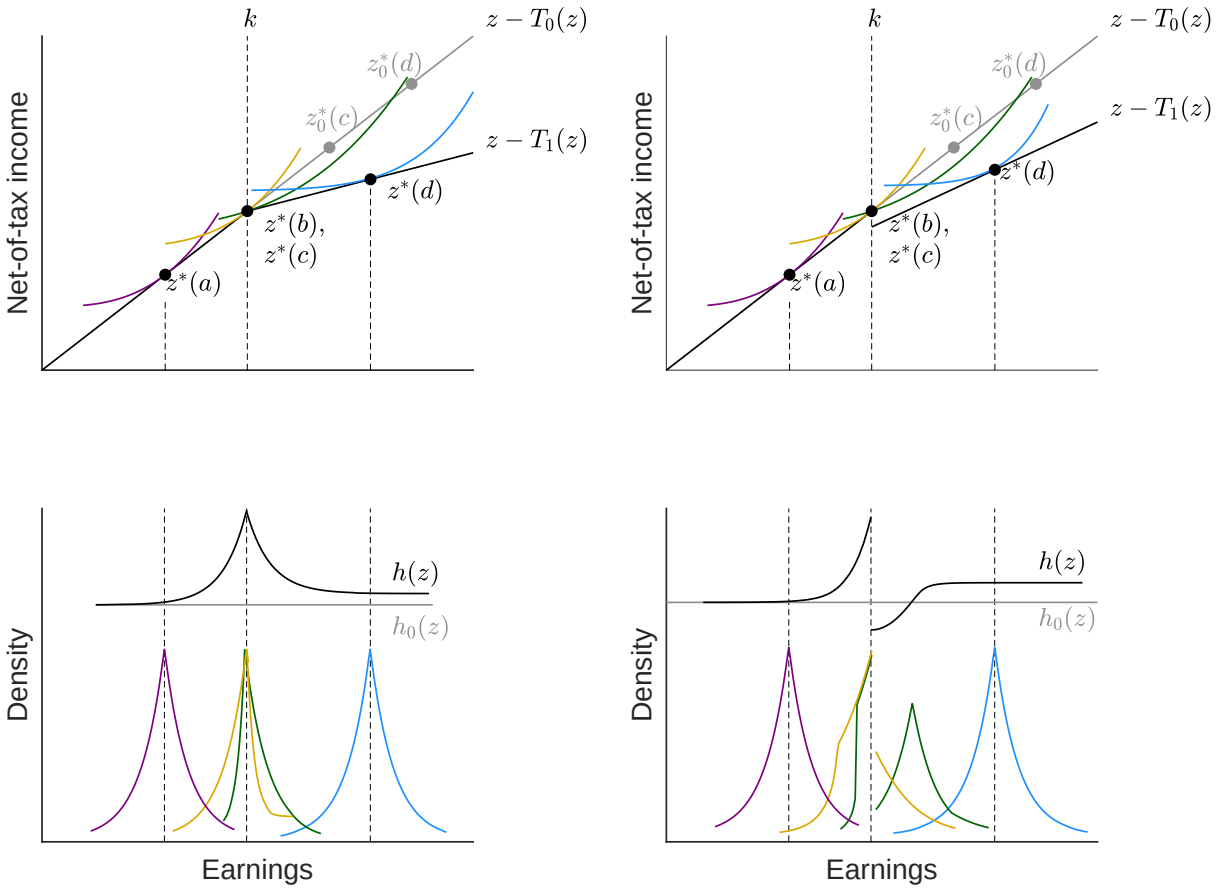
Panels (a) and (c) illustrate the construction of the indirect utility function around a progressive tax kink for the marginal non-buncher (type b). Panel (a) shows the taxpayer's budget constraint, plotted as a solid line, where $T_0(z)$ and $T_1(z)$ are the linear income taxes below and above the bracket threshold k, respectively. Panel (c) plots the indirect utility functions $v_0(z|b)$ and $v_1(z|b)$, which would be obtained if the linear tax functions $T_0(z)$ or $T_1(z)$ applied across all incomes. Type b's indirect utility function under the kinked tax schedule, plotted as a solid line, is given by $v_0(z|b)$ below k and $v_1(z|b)$ above k. Panels (b) and (d) show analogous illustrations for the marginal buncher (type c). This taxpayer's optimal frictionless income choices under the linear taxes $T_0(z)$ and $T_1(z)$ are denoted $z_0^*(c)$ and $z_1^*(c)$.

Figure 5: Type-conditional income density around a kink



This figure illustrates how the indirect utility functions from Figure 4 are used to compute the type-conditional income densities. The panels show the calculation for the marginal non-buncher (Panel (a)) and the marginal buncher (Panel (b)). Each panel illustrates the calculation of the type-conditional income density $g(z|n)$ at a (different) income $\tilde{z}$. We first identify the range of incomes that dominate $\tilde{z}$ for each taxpayer, corresponding to the horizontal dashed line, and we proceed as in Figure 3. The type-conditional densities are plotted in purple. For reference, the type-conditional density under the counterfactual linear taxes $T_0(z)$ and $T_1(z)$ are plotted in blue and in red, respectively.

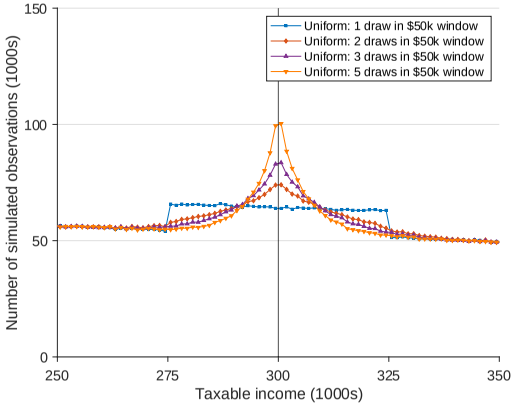
Figure 6: Aggregating type-conditional densities into an observable income density



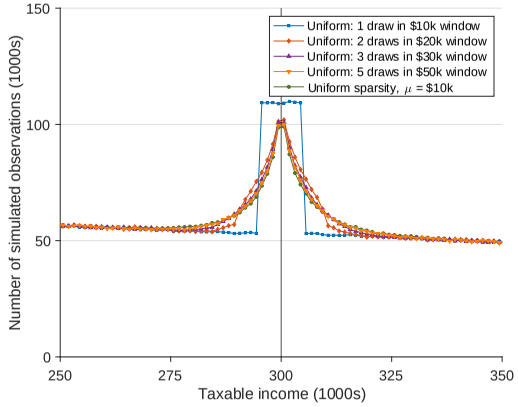
The top left panel of this figure shows the optimal frictionless income choice for agents of types a, b, c, and d in the presence of a kink at k, with each type's maximal indifference curve plotted in a different color. The panel below it illustrates the type-conditional income densities in corresponding colors. Summing across the type-conditional densities of these and the continuum of intervening types produces the observed income density, $h(z)$, which exhibits diffuse bunching around the bracket threshold. The counterfactual income density $h_0(z)$, which would be observed under the linear tax function $T_0(z)$, is plotted in gray for reference. The top right panel shows the analogous construction for a tax notch, with the resulting observed income density plotted below.

Figure 7: Simulated bunching patterns with sparsity-based frictions

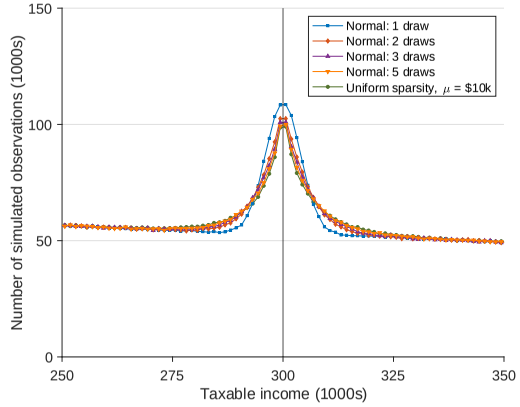
(a) Uniform draws, fixed support (kink)



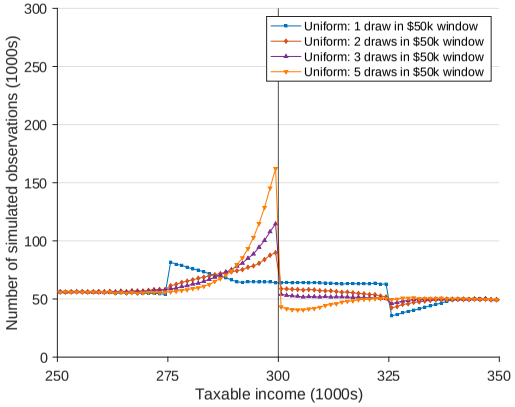
(b) Uniform draws, fixed density (kink)



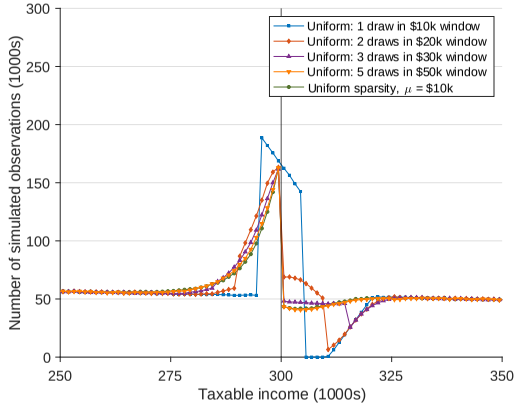
(c) Normal draws, fixed density (kink)



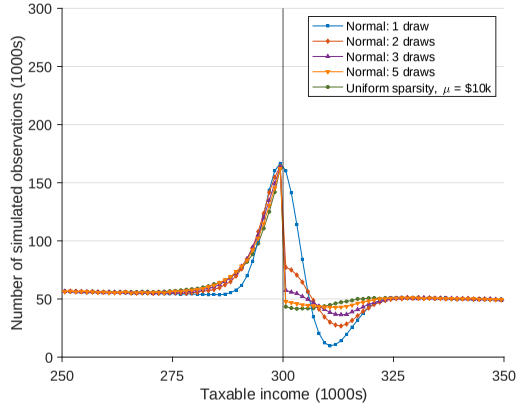
(d) Uniform draws, fixed support (notch)



(e) Uniform draws, fixed density (notch)



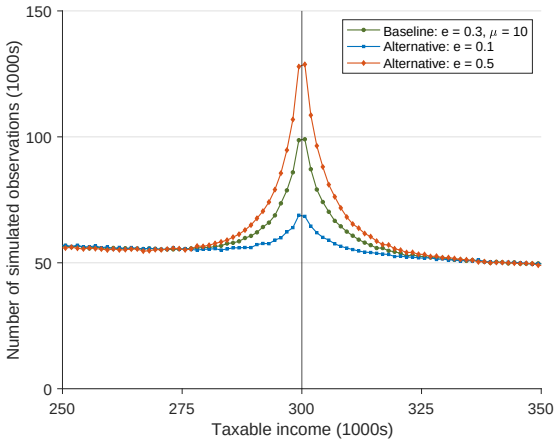
(f) Normal draws, fixed density (notch)



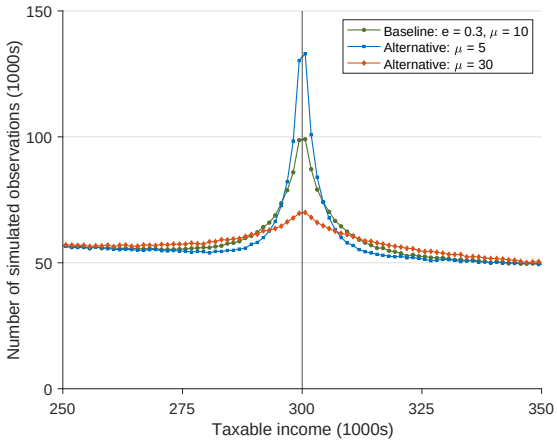
This figure plots simulated income densities around a bracket threshold under a model of frictions in which each taxpayer faces a sparse set of income opportunities drawn from around their preferred frictionless ("target") income. In all simulations, the marginal tax rate rises from $t_0 = 0.1$ to $t_1 = 0.2$ at the bracket threshold of $300,000. In Panels (d)–(f), the *level* of tax liability increases by $1000 at the bracket threshold. In Panels (a) and (d), each taxpayer chooses from M income opportunities drawn from a uniform distribution of width $50,000 around their target income, for $M = 1, 2, 3$, and 5. In Panels (b) and (e), each taxpayer chooses from M income opportunities drawn from a uniform distribution whose width is adjusted to hold fixed the density of opportunities around the target income. These panels also plot the "uniform sparsity model"—the limiting case as $M \to \infty$—in which income opportunities are a Poisson process with the same density of income opportunity draws at the target income. Panels (c) and (f) plot simulations in which income opportunities are drawn from a *normal* distribution centered at taxpayers' target incomes, with variance adjusted so that the density of opportunities is the same as in Panels (b) and (e).

Figure 8: Simulated effects of parameter variations on income densities

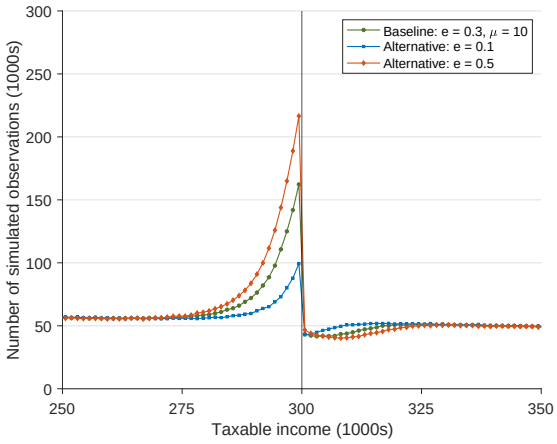
(a) Different elasticities (kink)



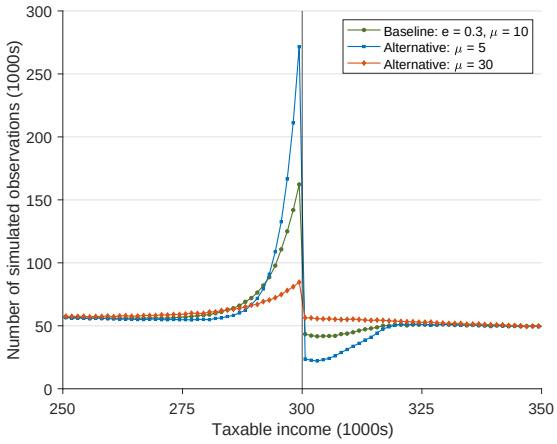
(b) Different lumpiness parameters (kink)



(c) Different elasticities (notch)



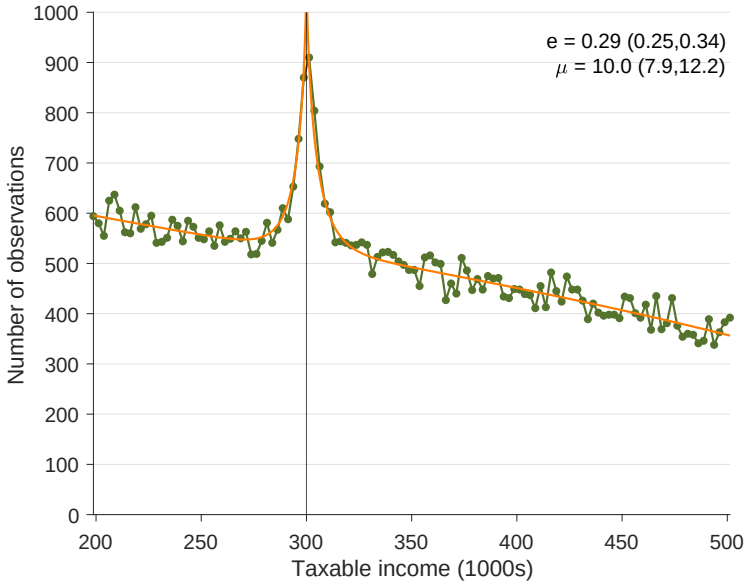
(d) Different lumpiness parameters (notch)



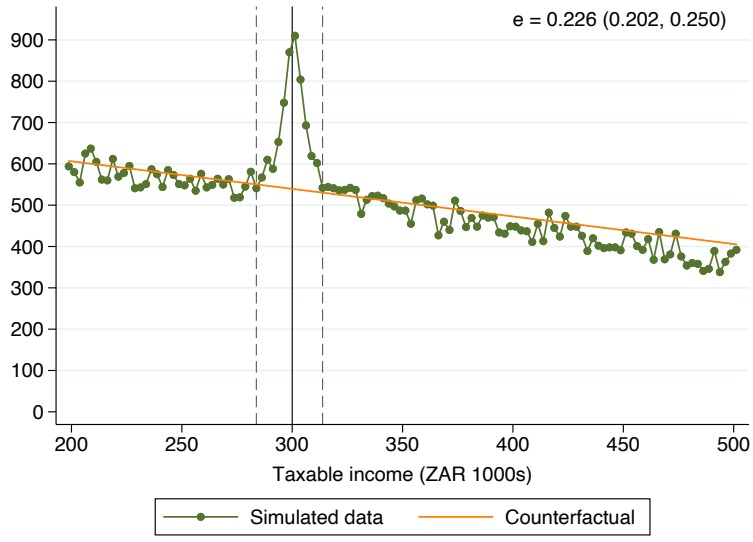
This figure plots income histograms from simulated data sets under the uniform sparsity model of frictions. For each simulation, we draw agents from an ability distribution with a linear density. We assume agents have a homogeneous income elasticity, e_0, and for each agent we then draw a sparse set of income opportunities from a Poisson process with a specified lumpiness parameter, μ_0. Each agent chooses the income opportunity that delivers the highest utility. We bin the resulting incomes to construct the income histograms displayed above. Panels (a) and (b) display simulated income histograms around a progressive tax kink for different values of e_0 and μ_0, respectively. Panels (c) and (d) display histograms around a tax notch. In each case, the marginal tax rate rises from 0.1 to 0.2 at \$300,000, and for the notch simulations in in Panels (c) and (d), the level of tax liability increases by \$1000.

Figure 9: Parameter estimates from simulated data

(a) Sparsity-based frictions estimator for a single simulation round



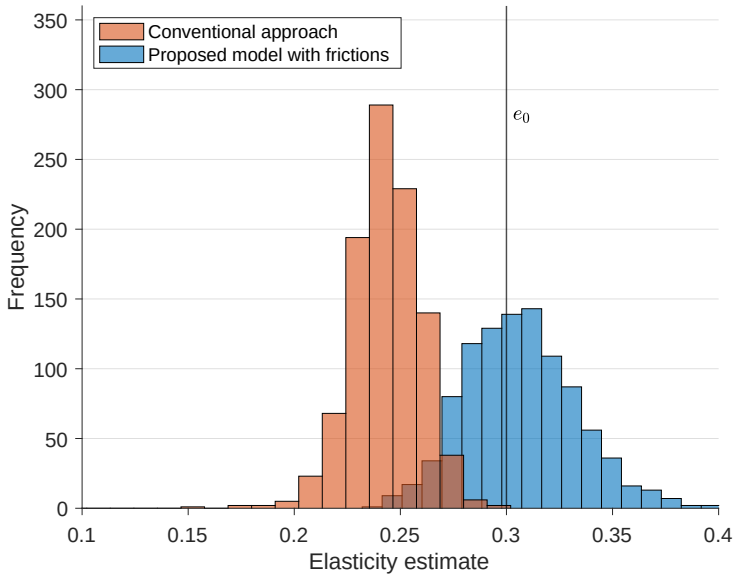
(b) Conventional bunching estimator for a single simulation round



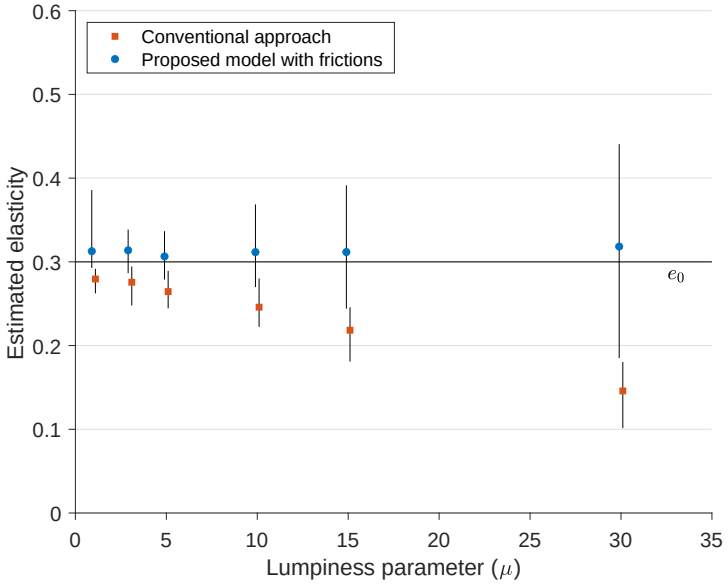
This figure displays the estimation of the maximum likelihood model and the conventional bunching estimator for one round of simulated data. The simulation is constructed as in Figure 8, with a true elasticity of $e_0 = 0.3$ and a lumpiness parameter of $\mu_0 = 10$, but using a smaller number of drawn observations ($M = 100{,}000$) to produce a level of sampling noise similar to that in our empirical application in Section 4. Panel (a) displays the results of applying the our maximum likelihood estimation method described in Section 2.5. Estimates of $\hat{e}$ and $\hat{\mu}$ and their 95 percent confidence intervals are reported in the upper corner. Panel (b) illustrates the conventional bunching estimator, applied to the same round of simulated data, resulting in an elasticity estimate well below the true value $e_0 = 0.3$. The vertical dashed lines display the algorithmically selected bunching window, and the orange line plots the best-fit polynomial to the data points outside the bunching window.

Figure 10: Elasticity estimates using the conventional approach

(a) Distribution of elasticity estimates under each approach

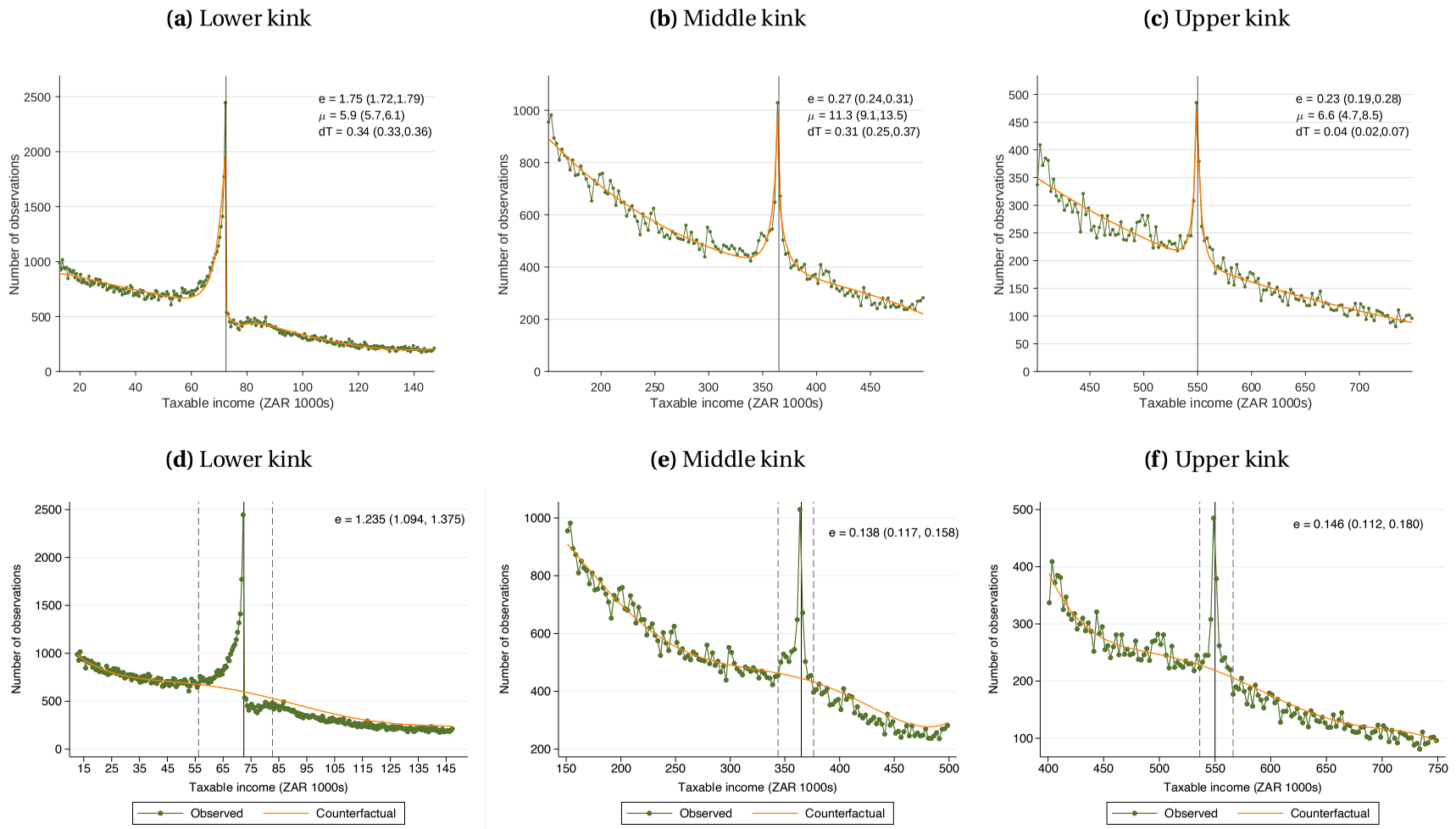


(b) Elasticity estimates under each approach for different lumpiness parameters



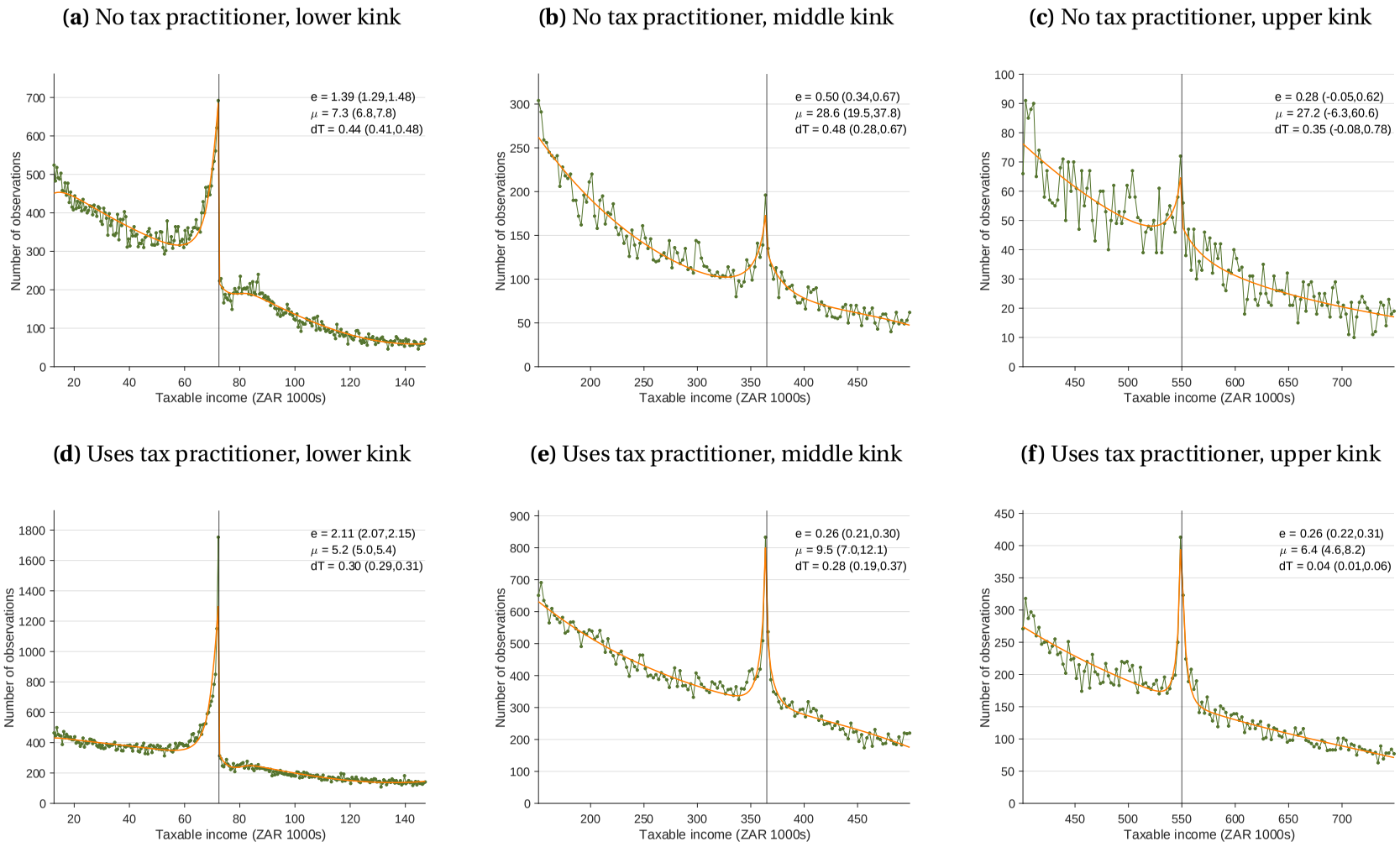
Panel (a) plots the histogram of elasticity estimates under the conventional approach (orange) and the maximum likelihood method allowing for frictions (blue). The vertical line at e_0 locates the true elasticity of the data generating process used to construct the simulated data sets. To construct Panel (b), we produce histograms like those in Panel (a) using simulated data with several different lumpiness parameters, holding fixed the true elasticity. Panel (b) displays the mean and 95 percent confidence intervals for the distribution of elasticity estimates in each case.

Figure 11: Bunching estimation with vs. without sparsity-based frictions: small business income taxes



Green points plot the income histograms of South African Small Businesses. In Panels (a)–(c), orange lines plot the best-fit income density under the uniform sparsity model. Figures report parameter estimates e (elasticity of taxable income), μ (average distance between income opportunities in ZAR 1000s), and dT (the estimated "as-if" discrete change in tax liability at the bracket threshold, in ZAR 1000s). Panels (d)–(f) present the conventional bunching estimator. Orange lines plot the estimated polynomial counterfactual density following the approach of Chetty et al. (2011). We choose the order of the polynomial and the bunching region (indicated by dashed lines) using the approach described in Appendix D. Numbers in parentheses indicate the 95 percent confidence interval on parameter estimates generated by the MLE method in Panels (a)–(c), and by bootstrapping in Panels (d)–(f).

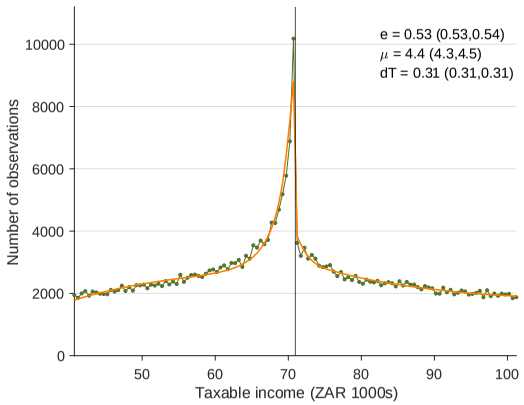
Figure 12: Firm heterogeneity by tax practitioner usage



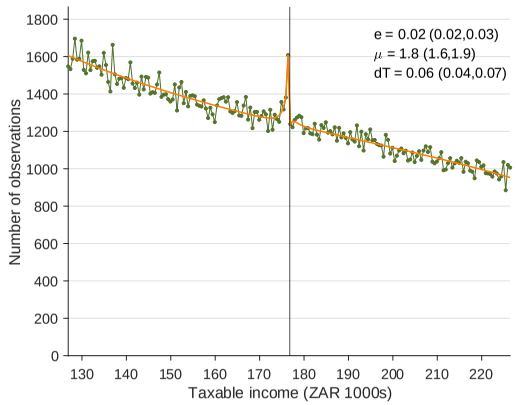
The format of these plots is the same as in Panels (a)–(c) of Figure 11, but the sample is split into firms that do and do not use professional tax practitioners to prepare their tax returns.

Figure 13: Model estimation: personal income taxes (self-employed individuals)

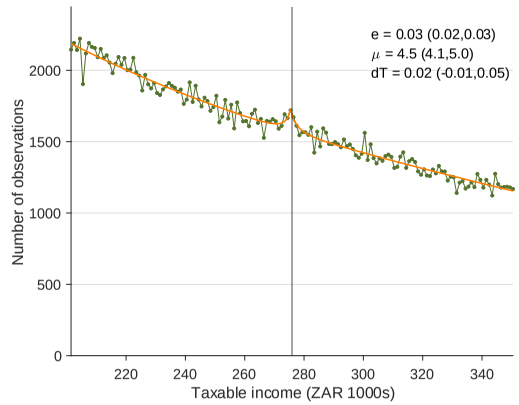
(a) First kink



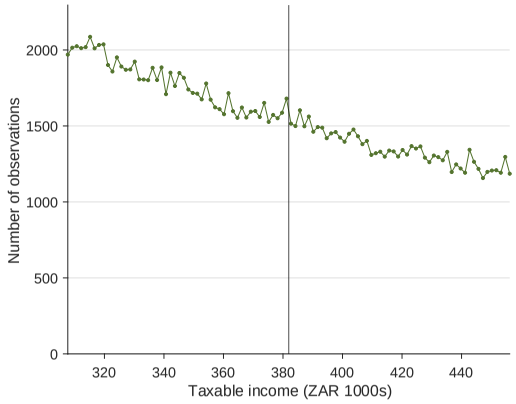
(b) Second kink



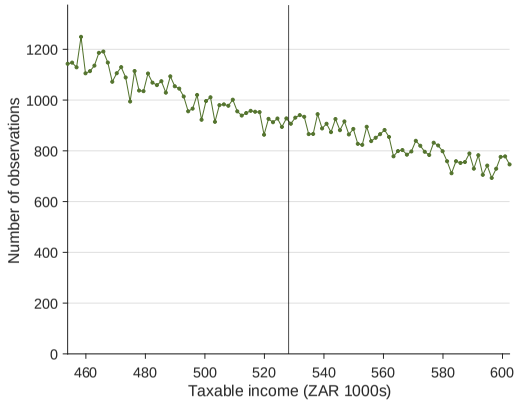
(c) Third kink



(d) Fourth kink



(e) Fifth kink



Green points plot the empirical histogram of individuals who report income from self-employment with different earnings in the data and orange lines plots the predicted density generated by the maximum likelihood estimation of the uniform sparsity model parameters e (elasticity of taxable income), μ (average distance between income opportunities in ZAR 1000s), and dT (the estimated "as-if" discrete change in tax liability at the bracket threshold, in ZAR 1000s). Numbers in parentheses indicate the 95 percent confidence interval on parameter estimates generated by the MLE method.

Table 1: Estimated bunching parameters: small business income tax

(a) Parameter estimates from model with frictions

	Elasticity of taxable income (e)		
	Lower	Middle	Upper
Full population	1.75 (1.72, 1.79)	0.27 (0.24, 0.31)	0.23 (0.19, 0.28)
Without tax practitioner	1.39 (1.29, 1.48)	0.50 (0.34, 0.67)	0.28 (-0.05, 0.62)
With tax practitioner	2.11 (2.07, 2.15)	0.26 (0.21, 0.30)	0.26 (0.22, 0.31)
	Lumpiness parameter (μ), in ZAR 1000s		
	Lower	Middle	Upper
Full population	5.9 (5.7, 6.1)	11.3 (9.1, 13.5)	6.6 (4.7, 8.5)
Without tax practitioner	7.3 (6.8, 7.8)	28.6 (19.5, 37.8)	27.2 (-6.3, 60.6)
With tax practitioner	5.2 (5.0, 5.4)	9.5 (7.0, 12.1)	6.4 (4.6, 8.2)
	As-if notch value (dT), in ZAR 1000s		
	Lower	Middle	Upper
Full population	0.34 (0.33, 0.36)	0.31 (0.25, 0.37)	0.04 (0.02, 0.07)
Without tax practitioner	0.44 (0.41, 0.48)	0.48 (0.28, 0.67)	0.35 (-0.08, 0.78)
With tax practitioner	0.30 (0.29, 0.31)	0.28 (0.19, 0.37)	0.04 (0.01, 0.06)

(b) Parameter estimates from conventional bunching estimator

	Elasticity of taxable income (e)		
	Lower	Middle	Upper
Full population	1.23 (1.14, 1.33)	0.14 (0.12, 0.16)	0.15 (0.11, 0.18)
Without tax practitioner	0.76 (0.66, 0.87)	0.11 (0.08, 0.15)	0.10 (0.06, 0.15)
With tax practitioner	1.51 (1.37, 1.65)	0.13 (0.11, 0.15)	0.14 (0.11, 0.17)

Panel (a) reports our maximum likelihood estimates of the elasticity of taxable income (e), the average distance between income adjustment opportunities (μ) and the revealed preference ("as-if") value of the change in tax liability at each bracket threshold. The values of μ and dT are measured in ZAR 1000s. Results are reported separately for the aggregate population, and for the subset of firms who do and do not use paid tax practitioners to prepare their tax returns. Panel (b) reports the estimated elasticity (e) from the conventional bunching estimator, using the method based on Chetty et al. (2011) and described in Appendix D.

Online Appendix

Diffuse Bunching with Frictions: Theory and Estimation

Santosh Anagol, Allan Davids, Benjamin B. Lockwood, Tarun Ramadorai

A Details of Bunching in the Presence of a Tax Notch

Figures A1–A3 illustrate details of the bunching model around kinks and notches in the frictionless model and with sparsity-based frictions.

Figure A1 illustrates bunching patterns that arise from a kink (panels a and b) or a notch (panels c and d) in a model without income frictions. The left panels a and c illustrate the choices of individual types of taxpayers, each of whom is depicted by a discrete dot. The right panels translate this discrete choice behavior into income densities with a continuum of types. In the presence of a kink, the frictionless model predicts an atom of mass at the bracket threshold with smooth densities on either side, generally with a discontinuity at the threshold due to the leftward shift and income compression of taxpayers who face the higher marginal tax rate above the threshold. In the presence of a notch, the model again predicts an atom of mass at the threshold and a smooth density to the left, but with an absence of mass (density equal to zero) in a dominated region of incomes just above the bracket threshold.

Figures A2 and A3 are analogous to Figures 4 and 5 in the paper, but in the presence of a notch rather than a kink. The notch in the budget constraint produces a discontinuity in the indirect utility functions plotted in Figures A2c and A2d. For illustrative purposes, type c is selected to be the type that is just indifferent between two levels of income, k and their optimal income choice under $T_1(z)$. (This is distinct from Figure 6 in the paper, where type c strictly prefers k.) As illustrated in Figure A3, the logic of Proposition 1 again carries through—the type-conditional density scales with the probability of drawing an income opportunity inside the shaded region of dominating incomes—although the notch sometimes produces a set of dominating income $\Theta(z|n)$ consisting of disjoint intervals, as shown in A3b. Under sparsity-based frictions, type-conditional densities exhibit positive density even in the so-called "dominated region" to the right of the notch because for some taxpayers an income opportunity in that region may be preferable to all of their other opportunities.

B Proof of Proposition 3

The simulations of bunching around a kink in Figure 7, panels (a)–(c), share a common structure. Agents draw an opportunity set consisting of M opportunities

$$\{z_1, z_2, \ldots, z_M\} = \left\{z^* + \varepsilon_1, z^* + \varepsilon_2, \ldots, z^* + \varepsilon_M\right\}, \tag{26}$$

where the ε are iid random draws from a particular distribution—uniform in panels a and b, and normal in panel c—which we can more generally denote F_ε, with density f_ε. Note that the probability of drawing some particular income $\tilde{z}$ for some specific income opportunity, such as z_1, is $f_\varepsilon(\tilde{z}-z^*)$, and so the probability of drawing $\tilde{z}$ among any of the M opportunities is $Mf_\varepsilon(\tilde{z}-z^*)$. Throughout this proof, we often suppress the dependence on agent type n to simplify notation.

A key observation is that choice behavior is determined by the opportunities nearest to an agent's target, i.e, the lowest positive ε—because all higher income opportunities are dominated by that draw—and the highest negative ε–because all lower income opportunities are dominated by that draw). We can view these two draws as a pair of order statistics, of a sort. The ultimate income density will depend solely on the distribution of these order statistics. The number of draws M, and the underlying distribution from which opportunities are drawn (i.e., the distribution of F_ε) are consequential only through their effect on these order statistics.

The order statistics themselves are straightforward to compute. The probability that $z^* + \varepsilon_j$ is the agent's preferred opportunity is simply the probability of drawing $z^* + \varepsilon_j$—which is just $f_\varepsilon(\varepsilon_j)$—times the probability that $z^* + \varepsilon_j$ is the best available opportunity, i.e., the probability that every other drawn opportunity is less desirable.

For $\varepsilon_j > 0$, another income opportunity $z^* + \varepsilon_k$ is less desirable if either $\varepsilon_k > \varepsilon_j$ or $z^* + \varepsilon_k < \underline{Z}(z^* + \varepsilon_j)$, with $\underline{Z}(\cdot)$ (defined as in the text) indicating the utility-equivalent income to $z^* + \varepsilon_j$ below the target z^*. Conditioning on the agent's type and utility function, let $\underline{\phi}(\varepsilon)$ and $\overline{\phi}(\varepsilon)$ denote the functions that compute the lower and upper utility-equivalent disturbances from z^* that deliver the same utility as a given disturbance ε, so that $\underline{Z}(z^* + \varepsilon) - z^* = \underline{\phi}(\varepsilon)$ and $\overline{Z}(z^* + \varepsilon) - z^* = \overline{\phi}(\varepsilon)$ for all ε.

Therefore the order statistic of interest—the probability that $z^* + \varepsilon_j$ is the agent's preferred opportunity—is

$$Mf_\varepsilon\left(\varepsilon_j\right)\left[1 - \left(F_\varepsilon\left(\varepsilon_j\right) - F_\varepsilon\left(\underline{\phi}\left(\varepsilon_j\right)\right)\right)\right]^{M-1} \tag{27}$$

in the case of $\varepsilon_j > 0$, and it is

$$Mf_\varepsilon\left(\varepsilon_j\right)\left[1 - \left(F_\varepsilon\left(\overline{\phi}\left(\varepsilon_j\right)\right) - F_\varepsilon\left(\varepsilon_j\right)\right)\right]^{M-1} \tag{28}$$

in the case of $\varepsilon_j < 0$. Together, these equations characterize the *type-conditional income density* at a given income, which we can write as

$$g\left(z^*(n)+x|n\right) := Mf_\varepsilon\left(\varepsilon_j\right)\left[1-\left(F_\varepsilon\left(\overline{\phi}(x)\right)-F_\varepsilon\left(\underline{\phi}(x)\right)\right)\right]^{M-1}. \tag{29}$$

This characterizes the probability that agents of type n earn income $z^*(n)+x$, where x is the distance above or below n's target income. We have used the fact that $\underline{\phi}(\varepsilon_j) = \varepsilon_j$ when $\varepsilon_j < 0$ and $\overline{\phi}(\varepsilon_j) = \varepsilon_j$ when $\varepsilon_j > 0$ by construction to combine the preceding equations.

Evaluating equation (29) at $x = 0$ gives $g(z^*(n)|n) = Mf_\varepsilon(0)$, the local density of opportunity draws around the agent's target income z^*. Section 2.2 provides heuristic logic suggesting that this local density is a sort of "sufficient statistic" for the income opportunity process, in the sense that a broad range of opportunity processes with the same local density of draws around the target income are approximated by the same type-conditional income density. Figures 7b and 7c illustrate this phenomenon for two parametric forms of the distribution F_ε (uniform and normal) and then adjusting that distribution to hold fixed the opportunity density $Mf_\varepsilon(0)$ while raising M, from which we observe apparent convergence as $M \to \infty$.

Here, we formalize and generalize that argument. We allow F_ε to be any distribution with a positive continuous density around zero, so that the probability of drawing income opportunities around the agent's income target is positive. We then define the following transformations of F_ε and f_ε:

$$F_\varepsilon^M(x) := F_\varepsilon\left(\frac{x}{M}\right), \tag{30}$$

with density

$$f_\varepsilon^M(x) := \frac{1}{M} f_\varepsilon\left(\frac{x}{M}\right). \tag{31}$$

The series of transformations densities underlying the simulations in Figure 7b and 7c are special cases of this general transformation. The transformation is constructed to adjust the spread of the distribution F_ε in a way that holds fixed the local density of opportunities around the target income, $Mf_\varepsilon(0)$. The transformation also preserves $F_\varepsilon^M(0)$, meaning that the target income remains at the same quantile in the distribution from which income opportunities are drawn. We call $F_\varepsilon^1(x) = F_\varepsilon(x)$ the *unitary distribution*. In keeping with the lumpiness parameter we define in the text, we will call this constant value $f_\varepsilon = \lambda = 1/\mu$.

Regarding F_ε^M and f_ε^M as any particular distributions from which M opportunities are drawn, equation (29) provides the type-conditional income density, which we now index by M:

$$g^M\left(z^*(n)+x|n\right) := Mf_\varepsilon^M(x)\left[1-\left(F_\varepsilon^M\left(\overline{\phi}(x)\right)-F_\varepsilon^M\left(\underline{\phi}(x)\right)\right)\right]^{M-1}. \tag{32}$$

Substituting the definitions from equations (30) and (31), this equation can be rewritten in

terms of the unitary distribution:

$$g^M\left(z^*(n)+x|n\right) = f_\varepsilon\left(\frac{x}{M}\right)\left[1-\left(F_\varepsilon\left(\frac{\overline{\phi}(x)}{M}\right)-F_\varepsilon\left(\frac{\underline{\phi}(x)}{M}\right)\right)\right]^{M-1}. \tag{33}$$

We are interested in the limit of this function $g^M(z|n)$ as $M\to\infty$. Employing the assumption that $F_\varepsilon(x)$ is differentiable at $x=0$, the definition of the derivative implies

$$\lim_{x\to 0, y\to 0} F_\varepsilon(x)-F_\varepsilon(y) = f_\varepsilon(0)(x-y) \text{ for } x\neq y.$$

Therefore letting $x=\overline{\phi}(x)/M$ and $y=\underline{\phi}(x)/M$ and noting $f_\varepsilon(0)=\lambda$, we have

$$F_\varepsilon\left(\frac{\overline{\phi}(x)}{M}\right)-F_\varepsilon\left(\frac{\underline{\phi}(x)}{M}\right) = \lambda\cdot\frac{\overline{\phi}(x)-\underline{\phi}(x)}{M}.$$

Substituting this result into equation (33) gives

$$g^\infty\left(z^*(n)+x|n\right) = \lambda\left[1-\lambda\cdot\frac{\overline{\phi}(x)-\underline{\phi}(x)}{M}\right]^{M-1} \tag{34}$$

$$= \lambda\exp\left[-\lambda\left(\overline{\phi}(x)-\underline{\phi}(x)\right)\right], \tag{35}$$

where the final line follows from $\lim_{M\to\infty}(1-rx/M)^M = e^{-rx}$, the definition of continuous exponential decay.

Using the definitions of $\underline{\phi}$ and $\overline{\phi}$, we can write this as the type-conditional density among taxpayers of type n as a function of income $\tilde{z}$:

$$g^\infty(\tilde{z}|n) = \lambda\exp\left[-\lambda\cdot\left(\overline{Z}(\tilde{z})-\underline{Z}(\tilde{z})\right)\right]. \tag{36}$$

Noting that in this setting the set of dominating incomes, $\Theta(\tilde{z}|n)$, is the interval $\left[\underline{Z}(\tilde{z}),\overline{Z}(\tilde{z})\right)$, and thus the measure of this set is $|\Theta(\tilde{z}|n)| = \overline{Z}(\tilde{z})-\underline{Z}(\tilde{z})$. Substituting this into equation (36) produces the type-conditional density of the uniform sparsity model as in Proposition 2. This demonstrates that any income opportunity process with continouous positive density around the target income converges to the uniform sparsity model with $\lambda = f_\varepsilon(0)$ as $M\to\infty$, proving Proposition 3.

C Separate Identification of Income Elasticity and Lumpiness Parameter

Here we discuss the separate identification of the income elasticity e and the lumpiness parameter μ, both numerically using simulations and analytically in the form of an analytic proof.

Figure A5a plots the joint distribution of estimates $(\hat{e}, \hat{\mu})$ for the 1000 simulation rounds underlying Figure 9a. The histogram of each marginal distribution is displayed outside of each axis. A number of notable features emerge. First, for both $\hat{e}$ and $\hat{\mu}$, the distribution of estimates is centered around the true parameter value. Averaging across simulation rounds, and the average value of $\hat{\mu}$ is 10.1 (measured in 1000s), close to the true values of $e_0 = 0.3$ and $\mu_0 = 10$.

Second, the spread of both distributions provides an indication of sampling error. In each round of simulated data, the maximum likelihood estimation procedure also provides a standard error estimate, and so a key question is whether this estimate gives an accurate picture of the degree of precision in the estimate. To explore this, we can compare the standard deviation of the distribution of $\hat{e}$ estimates, which is 0.026, to the average *estimate* of the standard error, which is 0.026, indicating that the maximum likelihood estimate of the standard error provides a good sense of the true degree of sampling uncertainty. In the case of μ, the standard deviation of the distribution of $\hat{\mu}$ is 1.121, and the average value of the estimated standard error is 1.099.

A third notable feature of Figure 9a is the upward slope in the cluster of joint estimates. This indicates that when $\hat{e}$ is overestimated due to sampling bias, it is likely that $\hat{\mu}$ is overestimated as well. To explore this phenomenon, Figure 9b plots model-generated income densities for five combinations of (e, μ). The thick solid line plots the baseline density with $e = 0.3$ and $\mu = 10$. The other four lines correspond to the (e, μ) pairs corresponding to the four square-shaped points in Figure 9b.

In Figure 9b, the densities corresponding to the points to the northwest and southeast of the baseline are easy to visually distinguish from the baseline, exhibiting substantially lower and higher densities at the kink, respectively. The reason for this pattern can be understood from the simulated densities in Figure 8. A higher elasticity e increases the density at the kink point by raising the total amount of bunching mass (Figure 8a). A *lower* value of the lumpiness parameter also increases the density at the kink point, by concentrating the excess mass more tightly around the kink (Figure 8b). Thus, the parameter combinations to the southeast of the baseline in Figure 9b correspond to densities with substantially higher density around the kink point, like the tallest density displayed in Figure 9b. The reverse is true for parameter combinations to the northwest of the baseline values, where the levels of both parameters

(low e and high μ) reinforce each other to push down the density at the kink. In contrast, parameter combinations to the northeast and southwest of the baseline have opposing effects on the density at the kink. They are still distinct, indicating that the model is identified, but their difference is more subtle, involving the density at intermediate points in between the kink point and the bounds of the income window. The pattern of points in Figure 9b corresponds to this visual impression: in the presence of sampling error, it is easier to distinguish—in a statistical sense—between data-generating processes with parameter pairs on the northwest-southeast axis than those on the northeast-southwest axis in Figure 9b.[33]

In sum, these points paint a clear picture of the performance of the maximum likelihood estimator when the model is correctly specified. Estimates of the elasticity and the lumpiness parameter appear consistent in that they are distributed around the true parameters of the data-generating process, and standard errors estimated by maximum likelihood are very close to the standard deviation of the distribution of estimates. They also highlight an important aspect of this model: estimation error in e and μ are likely to have the same sign. This result has important implications for the comparison of this model to the conventional elasticity estimator based on bunching mass.

We now show a related analytic result. Under the uniform sparsity model, for a given observed income density $h(z)$ and counterfactual density $h_0(z)$, if the target income response Δz (which identifies the elasticity e) and sparsity parameter λ are constant in a sufficiently wide region around the tax kink, then Δz and λ are separately identified.[34]

For this proof, we invoke the quadratic Taylor approximation from Assumption 1, which we maintain for the remainder of this section. As noted in Section 2.3, under this assumption each taxpayer's target income z^* (under a specified linear income tax) is a sufficient statistic for their type n, and thus we can index types by z^* rather than n. Therefore we will derive all results below in terms of the distribution of target incomes, which we denote $H_i^*(z)$, with density $h_i^*(z)$, for a given linear tax $T_i(z)$.

We now show how the counterfactual income densities $h_0(z)$ and $h_1(z)$ are related by the following lemma. This relationship must be shown—rather than following immediately from the definition of Δz—because although the income response Δz determines the shift of the *target* income distribution due to the tax reform, the *realized* income distribution differs from the target distribution due to frictions.

[33] Put differently, the estimator that we propose would find it easier to distinguish between data generated from "low e, high μ" and "low μ, high e" combinations than between "low μ, low e" and "high μ, high e" combinations.

[34] Note that our simulations in the paper assume a constant elasticity e, whereas for analytic simplicity this proof assumes a constant target income response Δz in levels. The two assumptions are quantitatively similar provided the region of bunching is small relative to the level of income at the kink k.

Lemma 1. *With constant target income response Δz in the vicinity of k, the counterfactual income CDFs and densities satisfy*

$$H_0(z) = H_1(z - \Delta z) \tag{37}$$

and

$$h_0(z) = h_1(z - \Delta z). \tag{38}$$

Proof. From equation (13)—though written using target incomes to index types—and Proposition 4, we can write the observed income CDF under $T_0(z)$ as

$$H_0(z) = \int_{\tilde{z}=-\infty}^{z} h_0(\tilde{z})d\tilde{z} = \int_{\tilde{z}=-\infty}^{z}\int_{z_0^*=-\infty}^{\infty} \lambda\left[-\lambda\cdot 2\left|\tilde{z}-z_0^*\right|\right] h_0^*(z_0^*)dz_0^*d\tilde{z}. \tag{39}$$

Similarly we can write

$$H_1(z) = \int_{\tilde{z}=-\infty}^{z} h_1(\tilde{z})d\tilde{z} = \int_{\tilde{z}=-\infty}^{z}\int_{z_1^*=-\infty}^{\infty} \lambda\left[-\lambda\cdot 2\left|\tilde{z}-z_1^*\right|\right] h_1^*(z_1^*)dz_1^*d\tilde{z}, \tag{40}$$

and therefore

$$H_1(z-\Delta z) = \int_{\tilde{z}=-\infty}^{z-\Delta z}\int_{z_1^*=-\infty}^{\infty} \lambda\left[-\lambda\cdot 2\left|\tilde{z}-z_1^*\right|\right] h_1^*(z_1^*)dz_1^*d\tilde{z}, \tag{41}$$

Using a change of variables with $\check{z} = \tilde{z} + \Delta z$ we have

$$H_1(z-\Delta z) = \int_{\check{z}=-\infty}^{z}\int_{z_1^*=-\infty}^{\infty} \lambda\left[-\lambda\cdot 2\left|\check{z}-\Delta z-z_1^*\right|\right] h_1^*(z_1^*)dz_1^*d\tilde{z}. \tag{42}$$

By the assumption of constant target income response, in the vicinity of the tax kink, *target incomes* under the linear tax $T_1(z)$ are lower than under $T_0(z)$ by Δz. That is, $z_1^* = z_0^* - \Delta z$. Substituting this into (42) gives

$$H_1(z-\Delta z) = \int_{\check{z}=-\infty}^{z}\int_{z_0^*=-\infty}^{\infty} \lambda\left[-\lambda\cdot 2\left|\check{z}-\Delta z-z_0^*+\Delta z\right|\right] h_1^*(z_0^*-\Delta z)dz_0^*d\tilde{z} \tag{43}$$

$$= \int_{\check{z}=-\infty}^{z}\int_{z_0^*=-\infty}^{\infty} \lambda\left[-\lambda\cdot 2\left|\check{z}-z_0^*\right|\right] h_1^*(z_0^*-\Delta z)dz_0^*d\tilde{z}. \tag{44}$$

Because target income rankings are preserved under a constant shift, $z_1^* = z_0^* - \Delta z$ implies

$$H_0^*(z_0^*) = H_1^*(z_1^*) = H_1^*(z_0^* - \Delta z). \tag{45}$$

and differentiating with respect to z_0^* gives

$$h_0^*(z_0^*) = h_1^*(z_0^* - \Delta z). \tag{46}$$

Substituting (46) into (44) gives the first equation in the lemma, (37), and differentiating (37) with respect to z gives (38), proving the lemma. □

We now define the "bunching mass function" $b(\epsilon)$ around the kink at k as the excess mass at income $k+\epsilon$ relative to the counterfactual density that would arise under the local linear tax function, i.e.,

$$b(\epsilon) = \begin{cases} h(k+\epsilon) - h_0(k+\epsilon) & \text{if } \epsilon < 0, \\ h(k+\epsilon) - h_1(k+\epsilon) & \text{if } \epsilon > 0. \end{cases} \tag{47}$$

In what follows, we prove two lemmas about the bunching mass function $b(\epsilon)$, which facilitate our numerical estimation and demonstrate that we can recover the target income response Δz for given distributions $H_0(z)$ and $H(z)$.

By construction, the integral of the bunching mass function is equal to the vertical distance between the income CDFs at k (Figure 1):

$$\begin{aligned} \int_{-\infty}^{\infty} b(\epsilon)\,d\epsilon &= \int_{-\infty}^{k} (h(z) - h_0(z))\,dz + \int_{k}^{\infty} (h(z) - h_1(z))\,dz \\ &= \int_{-\infty}^{k} h(z)dz - \int_{-\infty}^{k} h_0(z)dz + \int_{k}^{\infty} h(z)dz - \int_{k}^{\infty} h_1(z)dz \\ &= \underbrace{\int_{-\infty}^{\infty} h(z)dz}_{=1} - \underbrace{\int_{-\infty}^{k} h_0(z)dz}_{=H_0(k)} - \underbrace{\int_{k}^{\infty} h_1(z)dz}_{=1-H_1(k)} \\ &= H_1(k) - H_0(k) = B. \end{aligned} \tag{48}$$

Our next lemma characterizes the bunching mass function $b(\epsilon)$ explicitly. We focus on the case where $\epsilon > 0$, as the case where $\epsilon < 0$ can be handled symmetrically with identical techniques. Here it is useful to define some additional notation. Recall from Proposition 1 that $\Theta(z|n)$ denotes the set of incomes that utility-dominate $\tilde{z}$ for a taxpayer of type n:

$$\Theta(\tilde{z}|n) := \left\{ z \middle| u\left(z - T(z), z \middle| n\right) \geq u\left(\tilde{z} - T(\tilde{z}), \tilde{z} \middle| n\right) \right\}.$$

It is also useful to define the (weakly larger) set of incomes that *would* dominate $\tilde{z}$ under the counterfactual linear income tax $T_i(z)$ (where $i = 0$ or $i = 1$) which we denote $\Theta_i(\tilde{z}|n)$:

$$\Theta_i(\tilde{z}|n) := \left\{ z \middle| u\left(z - T_i(z), z \middle| n\right) \geq u\left(\tilde{z} - T_i(\tilde{z}), \tilde{z} \middle| n\right) \right\}. \tag{49}$$

Using equation (17) from the text, under Assumption 1 the measure of Θ_i takes a simple form:

$$\left|\Theta_i(\tilde{z}|n)\right| = 2\left|\tilde{z} - z_i^*(n)\right|. \tag{50}$$

Finally, for the purposes of the proof below we define the function $\delta(\epsilon, z_1^*)$ as the difference in the size of the dominating income set for a taxpayer targeting z_1^* under the kinked tax $T(z)$ relative to the linear tax $T_1(z)$:

$$\delta(\epsilon, z_1^*(n)) := \left|\Theta_1\left(k+\epsilon|n\right)\right| - \left|\Theta\left(k+\epsilon|n\right)\right|. \tag{51}$$

The definitions of Θ, Θ_1, and δ are illustrated in Figure A6. We can now characterize the bunching mass function $b(\epsilon)$ as follows.

Lemma 2. *The bunching mass function at income $k+\epsilon > k$ under the uniform sparsity model is*

$$b(\epsilon) = \int_{-\infty}^{k+\epsilon/2} \lambda \exp\left[-\lambda 2\left|k+\epsilon-z_1^*\right|\right]\left(\exp[\lambda\delta(\epsilon|z_1^*)]-1\right) h_1^*(z_1^*)\,dz_1^*. \tag{52}$$

Proof. Using equation (13) from the text, expressed using target incomes z_1^* as our index of types, we can write the observed income density at income $k+\epsilon$ as

$$\begin{aligned}
h(k+\epsilon) &= \int_{-\infty}^{\infty} g\left(k+\epsilon|z_1^*\right) h_1^*\left(z_1^*\right) dz_1^* && (53)\\
&= \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda\left|\Theta(k+\epsilon|z_1^*)\right|\right] h_1^*\left(z_1^*\right) dz_1^* && (54)\\
&= \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda\left(\left|\Theta_1(k+\epsilon|z_1^*)\right| - \delta(\epsilon, z_1^*)\right)\right] h_1^*\left(z_1^*\right) dz_1^* && (55)\\
&= \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda\left|\Theta_1(k+\epsilon|z_1^*)\right|\right] \exp\left[\lambda\delta(\epsilon, z_1^*)\right] h_1^*\left(z_1^*\right) dz_1^*, && (56)
\end{aligned}$$

where the second line substitutes the expression for the type-conditional density under uniform sparsity from Proposition 2, the third line uses the definition of $\delta(\epsilon, z_1^*)$ in (51), and the final line uses equation (50).

Similarly, we can express the counterfactual income density under the linear tax $T_1(z)$ as

$$h_1(k+\epsilon) = \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda\left|\Theta_1(k+\epsilon|z_1^*)\right|\right] h_1^*\left(z_1^*\right) dz_1^*. \tag{57}$$

Using the definition in (47), we find $b(\epsilon)$ by subtracting equation (57) from (56). Using the expression for $\left|\Theta_1(\tilde{z}|z_1^*)\right|$ in equation (50), this can be written as

$$b(\epsilon) = \int_{-\infty}^{\infty} \lambda \exp\left[-\lambda 2\left|k+\epsilon-z_1^*\right|\right]\left(\exp[\lambda\delta(\epsilon|z_1^*)]-1\right) h_1^*(z_1^*)\,dz_1^*. \tag{58}$$

Any taxpayer with a target income $z_1^* \geq k+\epsilon/2$ prefers income $k+\epsilon$ to any income below k. As a result, their set of incomes dominating $k+\epsilon$ is the same as the income set that would dominate $k+\epsilon$ under the counterfactual linear tax $T_1(z)$. That is, $\Theta(k+\epsilon|z_1^*) \equiv \Theta_1(k+\epsilon|z_1^*)$, and

thus $\delta(\epsilon|z_1^*) = 0$, for $z_1^* \geq k+\epsilon/2$. As a result, in equation (58) the integrand is zero for $z_1^* \geq k+\epsilon/2$, so this equation reduces to the one in the lemma. □

Lemmas 1 and 2 enable us to prove that under the uniform sparsity model, for a given counterfactual income distribution $H_0(z)$, an observed income distribution $H(z)$ can be consistent with only one target income response Δz. By construction of $b(\epsilon)$ (and equation 48), the integral $\int_{-\infty}^{\infty} b(\epsilon)d\epsilon$ is finite, and thus so is the portion in the positive domain, $\int_0^{\infty} b(\epsilon)d\epsilon$. By Lemma 2, the bunching mass function is strictly positive ($b(\epsilon) > 0$ for all ϵ). Together, these facts imply that the integral $\int_{\bar{\epsilon}}^{\infty} b(\epsilon)d\epsilon$ converges to zero as $\bar{\epsilon} \to \infty$. That is, for any desired precision $\varepsilon^* > 0$, there is a $\bar{\epsilon}$ such that $\int_{\bar{\epsilon}}^{\infty} b(\epsilon)d\epsilon < \varepsilon^*$. And from the definition of $b(\epsilon)$ we have

$$\begin{aligned}\int_{\bar{\epsilon}}^{\infty} b(\epsilon)d\epsilon &= \int_{\bar{\epsilon}}^{\infty} \left[h(k+\epsilon) - h_1(k+\epsilon)\right] d\epsilon \\ &= 1 - H\left(k+\bar{\epsilon}\right) - \left(1 - H_1\left(k+\bar{\epsilon}\right)\right) \\ &= H_1\left(k+\bar{\epsilon}\right) - H\left(k+\bar{\epsilon}\right). \end{aligned} \tag{59}$$

Thus by choosing $\bar{\epsilon}$ sufficiently large, we can ensure that $H_1(k+\bar{\epsilon}) - H(k+\bar{\epsilon}) < \varepsilon^*$, implying that the observed income distribution $H(z)$ approximates the counterfactual income distribution $H_1(z)$ to within ε^*. By assumption of our proposition, the target income response Δz is constant in a sufficiently wide region around the tax kink; we now exploit that assumption to assert that it is constant in a region that includes $k+\bar{\epsilon}$. By Lemma 1, we can thus write

$$H_1(k+\bar{\epsilon}-\Delta z) - H_0(k+\bar{\epsilon}) = 0 \implies \left|H\left(k+\bar{\epsilon}-\Delta z\right) - H_0\left(k+\bar{\epsilon}\right)\right| < \varepsilon^*. \tag{60}$$

For known $H(z)$ and $H_0(z)$, this equation can be used to identify Δz within an arbitrary degree of precision controlled by ε^*. This method depends on λ only through the speed at which $H(z)$ converges to $H_1(z)$ above k, and thus provided one chooses a sufficiently high $\bar{\epsilon}$ to achieve the desired degree of precision, variation in λ has an arbitrarily small effect on the estimated B, and thus on the estimated Δz. It is in this sense that Δz is separately identified from λ under the uniform sparsity model. The intuition for this result stems from the fact that the elasticity is identified entirely by the integral of the bunching mass function—B—which is not affected by the value of λ.

Having identified Δz, we have also identified $H_1(z)$ and $h_1(z)$. Our next lemma demonstrates how the sparsity parameter λ can be identified from the observed income density $h(z)$ and the target income response Δz.

Lemma 3. *Under the uniform sparsity model, the density at the kink satisfies*

$$h(k) = \lambda B + \int_{-\infty}^{k} h_0^*(z_0^*)\lambda \exp\left[-\lambda 2|k - z_0^*|\right] dz_0^* + \int_{k}^{\infty} h_1^*(z_1^*)\lambda \exp\left[-\lambda 2|k - z_1^*|\right] dz_1^*. \quad (61)$$

Proof. Any agent with a target income $z_0^* < k$ prefers income k to any $z > k$, regardless of whether a kink is present, and therefore the contribution of such agents to the observed income density $h(z)$ is equal to their contribution to the counterfactual density $h_0(z)$. Similarly, any agent with a target income $z_1^* > k$ prefers income k to any $z < k$, and therefore their contribution to $h(z)$ is equal to their contribution to $h_1(z)$.

Finally, for any agent with $z_1^* < k < z_0^*$, or equivalently, $k < z_0^* < k + \Delta z$, their preferred income under the kink is k, meaning the dominating income region is just a single value, k, implying that their type-conditional density at k is $\lambda \exp[-\lambda \cdot 0] = \lambda$. Combining these facts, the observed density at k can be written

$$h(k) = \int_{-\infty}^{k} h_0^*(z_0^*)\lambda \exp\left[-\lambda 2|k - z_0^*|\right] dz_0^* + \int_{k}^{\infty} h_1^*(z_1^*)\lambda \exp\left[-\lambda 2|k - z_1^*|\right] dz_1^* + \lambda \int_{k}^{k+\Delta z} h_0^*(z_0^*) dz_0^*. \quad (62)$$

Note that the integral in the final term is equal to $\lambda(H_0(k + \Delta z) - H_0(k))$, and by Lemma 1, this is equal to $\lambda(H_1(k) - H_0(k)) = \lambda B$. This proves the lemma. □

Next, we produce the following lemma, which demonstrates that when the observed income density under a linear tax can be written as a polynomial, then the density of *target incomes* under that tax can also be written as a polynomial, with a simple relationship between the coefficients of the two polynomials.

Lemma 4. *Under the uniform sparsity model, if the Taylor approximation around income $\tilde{z}$ of the observed income density under a given linear tax $T_i(z)$ has coefficients α_j, i.e.,*

$$h_i(z) \approx \alpha_0 + \alpha_1(z - \tilde{z}) + \frac{\alpha_2}{2}(z - \tilde{z})^2 + \ldots, \quad (63)$$

then the Taylor approximation of the (unobservable) density of target incomes, $h_i^*(z)$, around $\tilde{z}$ has coefficients α_j^*, where

$$\alpha_j^* = \alpha_j - \frac{\alpha_{j+2}}{(2\lambda)^2}. \quad (64)$$

Proof. Under the uniform sparsity model, the density of observed and target incomes are related by the following equation (the subscript i indexing the linear tax is suppressed for

readability, as the result turns out not to depend on the tax, provided that it is linear):

$$h(z) = \int_{z^*=-\infty}^{z} \lambda \exp\left[-2\lambda\left(z - z^*\right)\right] h^*(z^*) dz^* + \int_{z^*=z}^{\infty} \lambda \exp\left[-2\lambda\left(z^* - z\right)\right] h^*(z^*) dz^*. \quad (65)$$

Employing a change of variables in each integral, with $x = z - z^*$ (so $\frac{dx}{dz^*} = -1$) and $y = z^* - z$ (so $\frac{dy}{dz^*} = 1$), we can write

$$h(z) = \frac{1}{2}\left\{\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] h^*(z-x) dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] h^*(z+y) dy\right\}. \quad (66)$$

Intuitively, this equation states that the observed density at z is a weighted average of the target income density around z, where the weights decline exponentially (with parameter 2λ) as the target income gets farther from z in either direction.

This insight allows us to approximate the relationship between $h(z)$ and $h^*(z)$ around $\tilde{z}$ arbitrarily well using Taylor expansions. Let $\hat{h}^*(z)$ denote the Taylor expansion of the target income density $h^*(z)$ around income $\tilde{z}$:

$$\hat{h}^*(z) = \alpha_0^* + \alpha_1^*(z - \tilde{z}) + \frac{\alpha_2^*}{2}(z - \tilde{z})^2 + \ldots \quad (67)$$

where $\alpha_0^* = h^*(\tilde{z})$, $\alpha_1^* = h^{*\prime}(\tilde{z})$, $\alpha_2^* = h^{*\prime\prime}(\tilde{z})$, etc., with "$*$" superscripts indicating that these are the Taylor expansion coefficients of the *target* income density. Substituting this approximation into equation (66) evaluated at $z = \tilde{z}$, we have

$$h(\tilde{z}) = \frac{1}{2}\left\{\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right]\left(\alpha_0^* - \alpha_1^* x + \frac{\alpha_2^*}{2}x^2 - \frac{\alpha_3^*}{3!}x^3 + \frac{\alpha_4^*}{4!}x^4 + \ldots\right) dx + \right.$$
$$\left. \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right]\left(\alpha_0^* + \alpha_1^* y + \frac{\alpha_2^*}{2}y^2 + \frac{\alpha_3^*}{3!}y^3 + \frac{\alpha_4^*}{4!}y^4 + \ldots\right) dy\right\}. \quad (68)$$

Collecting terms by coefficients, we can rewrite this as

$$\begin{aligned}
h(\tilde{z}) = \; & \alpha_0^* \cdot \frac{1}{2}\left\{\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] dy\right\} \\
& + \alpha_1^* \cdot \frac{1}{2}\left\{-\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] y dy\right\} \\
& + \frac{\alpha_2^*}{2} \cdot \frac{1}{2}\left\{\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^2 dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] y^2 dy\right\} \\
& + \frac{\alpha_3^*}{3!} \cdot \frac{1}{2}\left\{-\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^3 dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] y^3 dy\right\} \\
& + \frac{\alpha_4^*}{4!} \cdot \frac{1}{2}\left\{\int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^4 dx + \int_{y=0}^{\infty} 2\lambda \exp\left[-2\lambda y\right] y^4 dy\right\} \\
& + \ldots
\end{aligned}$$

All terms multiplying odd-numbered Taylor coefficients (α_1^*, α_3^*, etc.) consist of two equal integrals with opposite signs, and they therefore cancel. The term multiplying α_0^* contains two identical integrals, each of which is an integral over an exponential density function, and therefore they each integrate to 1. More generally, terms multiplying even-numbered Taylor coefficients (α_0^*, α_2^*, etc.) consist of two equal integrals, and therefore their average is equal to either of them. Putting this together, we can simplify the above expression to

$$h(\tilde{z}) = \alpha_0^* + \frac{\alpha_2^*}{2!} \cdot \int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^2 dx + \frac{\alpha_4^*}{4!} \cdot \int_{x=0}^{\infty} 2\lambda \exp\left[-2\lambda x\right] x^4 dx + \ldots \tag{69}$$

We can now make use of a convenient property of exponential distributions: when x is exponentially distributed with rate parameter β, the expectation of x^m is $m!/\beta^m$. Each of the above integrals represents such an expectation, with x exponentially distributed with rate parameter 2λ. Therefore we can rewrite equation (69) as

$$\begin{aligned}
h(\tilde{z}) &= \alpha_0^* + \frac{\alpha_2^*}{2!} \cdot \frac{2!}{(2\lambda)^2} + \frac{\alpha_4^*}{4!} \cdot \frac{4!}{(2\lambda)^4} + \ldots \\
&= \alpha_0^* + \frac{\alpha_2^*}{(2\lambda)^2} + \frac{\alpha_4^*}{(2\lambda)^4} + \ldots \\
&= \alpha_0^* + \sum_{j=1}^{\infty} \frac{\alpha_{2j}^*}{(2\lambda)^{2j}}.
\end{aligned} \tag{70}$$

We do not observe the coefficients α_j^* because they are features of the unobserved distribution of target incomes. But we do observe features of $h(z)$. And the two sets of coefficients can be related using equation (70).

Letting $\hat{h}(z)$ denote the Taylor expansion around $\tilde{z}$ of the *observed* income density with

coefficients $\alpha_0 = h(\tilde{z})$, $\alpha_1 = h'(\tilde{z})$, $\alpha_2 = h''(\tilde{z})$, etc., we have, from equation (70),

$$\alpha_0 = h(\tilde{z}) = \alpha_0^* + \sum_{j=1}^{\infty} \frac{\alpha_{2j}^*}{(2\lambda)^{2j}}. \tag{71}$$

To find α_1, we differentiate equation (70) with respect to $\tilde{z}$, noting that the coefficients α_j^* are functions of $\tilde{z}$:

$$\alpha_1 = h'(\tilde{z}) = \alpha_0^{*\prime}(\tilde{z}) + \sum_{j=1}^{\infty} \frac{\alpha_{2j}^{*\prime}(\tilde{z})}{(2\lambda)^{2j}}. \tag{72}$$

By definition of the Taylor expansion, $\alpha_j^{*\prime}(\tilde{z}) = \alpha_{j+1}^*(\tilde{z})$, so we have

$$\alpha_1 = \alpha_1^* + \sum_{j=1}^{\infty} \frac{\alpha_{2j+1}^*}{(2\lambda)^{2j}}. \tag{73}$$

Similarly,

$$\begin{aligned} \alpha_2 &= \alpha_2^* + \sum_{j=1}^{\infty} \frac{\alpha_{2j+2}^*}{(2\lambda)^{2j}} \\ &= (2\lambda)^2 \left(\sum_{j=1}^{\infty} \frac{\alpha_{2j}^*}{(2\lambda)^{2j}} \right). \end{aligned} \tag{74}$$

Combining equations (71) and (74), we have

$$\alpha_0 - \frac{\alpha_2}{(2\lambda)^2} = \alpha_0^*. \tag{75}$$

By a similar process, it can be shown that

$$\alpha_1 - \frac{\alpha_3}{(2\lambda)^2} = \alpha_1^*, \tag{76}$$

and more generally,

$$\alpha_j - \frac{\alpha_{j+2}}{(2\lambda)^2} = \alpha_j^*. \tag{77}$$

This proves the lemma. □

With this lemma, using the known densities $h_0(z)$ and $h_1(z)$ we can approximate the target income densities $h_0^*(z)$ and $h_1^*(z)$ with an arbitrary degree of precision. We can then use Lemma 3 to compute λ.

This completes the proof that Δz and λ are separately identified under the uniform sparsity model.

D Details of the conventional kink-based bunching estimator

We apply the conventional bunching estimator based on Saez (2010) to estimate the income elasticity of the simulated data sets underlying Figure 9. We use as our baseline the implementation described in Chetty et al. (2011), which builds on Saez (2010) by estimating a counterfactual using a smoothed polynomial regression. Appendix F presents results using a number of alternative implementations of the conventional approach.

This estimation procedure involves two steps, first estimating a counterfactual income density based on the income density excluding data points near the kink, and then using the counterfactual density to estimate the excess mass from which the elasticity is recovered. To estimate the counterfactual density, we fit a polynomial of a specified degree to the observed income density, excluding the data in a specified window around the kink, using the following specification:

$$C_j = \sum_{i=0}^{q} \beta_i^0 \cdot (Z_j)^i + \sum_{i=R_l}^{R_u} \gamma_i^0 \cdot \mathbf{1}[Z_j = i] + \epsilon_j^0. \tag{78}$$

Here, q denotes the order of the polynomial, and R_l and R_u denote the lower and upper bounds of the "bunching window" near the kink, which is excluded from the polynomial estimation.[35] When estimating the polynomial regression, we follow Chetty et al. (2011) and impose an "integration constraint" such that the total count of observations across the empirical distribution equals the integral of observations under the counterfactual density across the plotted region.[36]

The second step is to compute the excess mass of incomes around the kink relative to this counterfactual density. Using equation (78), we compute the counterfactual mass in each bin within the bunching window, $\hat{C_j^0}$. Subtracting this predicted mass from the observed density yields the estimated excess number of individuals who report incomes near the kink relative to this counterfactual distribution:

$$\hat{B} = \sum_{i=R_l}^{R_u} C_j - \hat{C_j^0} = \sum_{i=R_l}^{R_u} \hat{\gamma}^0. \tag{79}$$

We then map this excess mass estimate to an estimated elasticity using the approximation from

[35] The convention in Chetty et al. (2011) is to set a symmetric bunching window, such that $R_l = -R_u$. We allow for the possibility of an asymmetric bunching window, following the approach in Bosch, Dekker and Strohmaier (2020), which we detail below.

[36] Kleven (2016) notes that imposing an integration constraint may bias the elasticity estimate: "This approach may introduce bias, especially in relatively flat distributions in which interior responses do not affect bin counts (except at the very top of the distribution away from the threshold being analyzed). It would be feasible to implement a conceptually more satisfying approach that does not have this potential bias, but for the reasons stated above, it will matter very little in most applications." As we discuss below and in Appendix F, our results confirm that the integration constraint introduces bias, and that the introduced bias may be substantial.

Chetty et al. (2011):

$$\hat{e} \approx \frac{\hat{B}}{z^* \cdot h_0(z^*) \cdot \log(\frac{1-t_0}{1-t_1})}. \tag{80}$$

Standard errors for $\hat{e}$ are estimated using a bootstrap procedure. We resample with replacement from the underlying distribution of firms 1000 times, re-estimating the elasticity each time, and defining the standard error as the standard deviation of the distribution of $\hat{e}$ estimates.

This conventional estimation method relies on three parameter inputs: the lower and upper bounds of the bunching window (R_l and R_u) and the order of the polynomial (q). These are often left to the discretion of the researcher to be chosen via "visual inspection." We instead follow the algorithmic approach proposed in Bosch, Dekker and Strohmaier (2020), which allows the polynomial order and the bunching region to be informed by the data itself.[37]

E Robustness to polynomial degree

Beginning with our simulation-based results, Figure A7 reproduces the estimations in Figure 9 assuming different polynomial degrees for the ability density (Panel (a)) or the counterfactual density outside the bunching window (Panel (b)). Our proposed method is not sensitive to misspecification in the polynomial order: the estimated elasticity is close to the true value of the data generating process, $e_0 = 0.3$, for each specification. In contrast, under the conventional bunching estimator, greater flexibility (i.e., a higher polynomial degree) causes the best fit of the counterfactual density to be "pulled upward" into the bunching mass, underestimating the bunching mass and producing a lower estimate of the elasticity. Each estimate also reports the Bayesian Information Criteria, which is minimized under the linear fit in each case.

[37] This approach proceeds in five steps: (1) Estimate equation (78) with no bunching window—so that the polynomial estimation excludes only the bins adjacent to the kink—for a range of polynomial orders, retaining the specification that minimizes the Bayesian Information Criterion (BIC). (2) Define the lower bound of the bunching window as the leftmost set of two adjacent bins below the threshold where the actual count in each bin exceeds the 95 percent confidence interval of the predicted bin counts from equation (78), and define the upper bound using an analogous procedure to the right of the kink. (3) Repeat steps (1) and (2), widening the bunching window by one bin above and below the kink each time. Each such iteration produces a candidate set of bounds for a bunching window. (4) From the resulting distributions of candidate bounds, choose the modal lower bound and upper bound to define the preferred bunching window. (5) Using this preferred bunching window, re-estimate the final counterfactual regression with the preferred polynomial order as in Step (1).

F Alternative specifications for the conventional approach

In Section 3, we compare the elasticities produced under our approach to the elasticity estimates produced under the approach developed in Chetty et al. (2011), one of the most widely used conventional bunching estimators. This approach involves fitting a flexible polynomial to the observed data, excluding the observations in the bunching region, and uses this to construct a single counterfactual which represents the counterfactual distribution that would occur if the lower tax rate below the kink threshold also applied above the threshold. As we discuss in the main text, this approach imposes an "integration constraint" such that the total integral of population across the empirical distribution equals the total integral under the counterfactual distribution. The integration constraint makes the assumption that all of the bunching mass comes from the income distribution in the underlying histogram and rules out the possibility that any mass shifts beyond the region depicted in the histogram. Given the counterfactual of Chetty et al. (2011) assumes that the lower tax rate below the kink applies above the kink, this means that the entirety of the bunching mass gets reallocated above the kink into the income bins depicted in the histogram. The practical implication of this is that the counterfactual density is shifted upward in order to ensure that the total integral of population across the empirical distribution equals the total integral under the counterfactual distribution.

In this section, we compare our estimator to other conventional bunching estimators that differ in how they construct counterfactuals, namely Saez (2010) and Mortenson and Whitten (2016), the working paper that preceded Mortenson and Whitten (2020). We illustrate the differences between these approaches in Figure A8a. The approach developed in Saez (2010) constructs two linear counterfactuals on either side of the kink with the assumption that the densities are uniformly distributed on either side of the threshold. In order to construct the counterfactual, the approach takes the average value of the densities that occur outside of the bunching window and extrapolates that density through to the kink threshold. This is done on either side of the kink resulting in two counterfactuals. An alternative approach is developed in Mortenson and Whitten (2016) who construct a piecewise linear counterfactual on either side of the kink, in a similar vein to Saez (2010), but to allow for that counterfactual to take into account the slope of the observed densities on either side of the kink. Finally, we also consider an implementation of the approach in Chetty et al. (2011) where we do not impose the "integration constraint." This allows for the possibility that the bunching mass may be reallocated to income bins beyond the region depicted in the histogram, which would cause the total integral under the counterfactual distribution to be smaller than the total integral of population across the empirical distribution. The practical implication of this is that the

counterfactual distribution is shifted downward relative to the approach which imposes the integration constraint, as is depicted in Figure A8a.

Next, we compare the elasticities produced under these four approaches to our estimates for varying values of the lumpiness parameter. We report these results in Figure A8b. Imposing the integration constraint in the Chetty et al. (2011) approach produces lower elasticities than when the constraint is not imposed. Intuitively, by imposing the constraint, the counterfactual density is shifted upward, which causes the estimate of bunching to fall, leading to a lower elasticity. The Mortenson and Whitten (2016) elasticities are very similar to the Chetty et al. (2011) elasticities without the integration constraint. In that sense, the specification is nearly identical to the counterfactual specification in Mortenson and Whitten (2016), apart from the fact that the latter approach estimates two counterfactuals on either side of the threshold, thereby allowing for a different slope on the counterfactual on either side of the kink threshold. The visual similarity between these counterfactuals is evident in Figure A8a. Out of all of the conventional approaches, the Saez (2010) approach produces the largest elasticities. The reason for this becomes evident when considering the counterfactuals produced in Figure A8a. Given the empirical distribution slopes downwards, by assuming uniformly distributed densities, the Saez (2010) approach produces a counterfactual that is significantly lower than the other counterfactuals in the bunching region above the kink, leading to a higher measure of bunching, and a higher estimated elasticity.

For smaller values of the lumpiness parameter, only Mortenson and Whitten (2016) and the Chetty et al. (2011) approach without the integration constraint can recover the true elasticity. However, for large values of the lumpiness parameter, not even these approaches are able to recover the true elasticity and exhibit a significant downward bias, whereas our approach can consistently recover the elasticity, irrespective of the extent of lumpiness in the observed empirical distribution.

G Comparison to the conventional notch-based bunching estimator

Kleven and Waseem (2013) (hereafter KW) proposes a method for estimating the elasticity of taxable income based on bunching around a notch in the tax schedule. Here we apply that method to simulated data with sparsity-based frictions. Figure A9 plots an income histogram for one simulation round with parameters identical to those in Figure 9, except in this case we impose a notch value of $1000.

Panel (a) displays the results of applying our maximum likelihood estimation with frictions.

This method delivers estimates of the elasticity and the lumpiness parameter that are close to the parameters of the data-generating process ($e_0 = 0.3$ and $\mu_0 = 10$), with confidence intervals that contain the true parameters.

Panel (b) displays the results of applying the KW notch-based estimator. In this model, the presence of taxpayers at incomes that are strictly dominated (i.e., at which post-tax income is lower *and* labor effort—as evidenced by pre-tax income—is higher than at k) is explained by assuming that a share a^* of taxpayers face frictions that render them unresponsive to the notch. According to this model, the structural or long-run elasticity is a function of the bunching mass that would be measured on a longer horizon when all taxpayers are responsive to the notch. Thus, the model proceeds by first estimating the bunching mass relative to the counterfactual frequency at k—denoted b in the figure—then scaling this estimate up by $1/(1-a^*)$ to adjust for under-responsiveness from frictions. The resulting rescaled mass is used to compute the structural elasticity.

Following KW, we compute the bunching mass b by visually specifying a lower bound z^L below which no excess bunching mass is apparent. We then compute an upper bound z^U by iteration to satisfy two conditions: the counterfactual frequency (plotted in orange in Panel (b)) is the quintic best-fit to the empirical histogram outside the excluded bunching window $[z_L, z_U]$, and the excess bunching mass in the interval $[z_L, k]$ fills the cumulative gap between the empirical histogram and the counterfactual frequency across the interval $[k, z_U]$.

Having identified the counterfactual frequency, we can compute a^*—the share of unresponsive taxpayers—as the ratio of the empirical histogram to the counterfactual density in the dominated income range.[38] This value is reported in Panel (b). Rescaling the bunching estimate b by $1/(1-a^*)$ and multiplying by the income bin width ($2,500 in these simulations) we get an estimate of the income change induced by the marginal buncher (Δz in the notation of KW) from which we compute the elasticity e using KW equation equation (5).[39]

The resulting elasticity estimate is $\hat{e} = 0.61$—a value that is higher than the elasticity of the data-generating process, $e_0 = 0.3$.[40] This overestimation comes from a misinterpretation of the of the presence of mass in the dominated income range. In the presence of sparsity-based friction, that mass is explained by the fact that some agents with an opportunity in the

[38] Under the tax parameters of this simulation, the upper bound of the dominated range z^D satisfies $k - T_0(k) = z^D - T_1(z^D)$, implying $z^D = 301{,}250$. This region is small relative to the tax systems considered in KW, which have larger implied notch values.

[39] KW uses alternative notation in which the piecewise-linear tax schedule is denoted $T(z) = t \cdot z + (\Delta T + \Delta t \cdot z) \cdot 1\{z > k\}$. Our simulated tax system with $T(z) = 0.1 \cdot z$ for $z \leq k$ and $T(z) = 31{,}000 + 0.2 \cdot (z - k)$ for $z \leq k$ implies values of $t = 0.1$, $\Delta t = 0.1$, and $\Delta T = -29{,}000$ using their notation.

[40] Re-running this estimation procedure on 1000 rounds of simulated data, we find an average estimated value of 0.78; the distribution of estimates has right skew, reflecting a right tail of high estimates that arise from simulation variation where the density in the dominated range is high, and thus the rescaling factor $1/(1-a^*)$ is large.

dominated region do not happen to draw any other opportunities that are more desirable; it is not an indication that some share of agents are unresponsive, despite having an available dominating option. This distinction can be seen sharply in Panels (c) and (d) of Figure 8, which plots simulated income distributions around a notch with alternative lumpiness parameters. When frictions increase, the income density in the dominated region becomes *higher* than the flat counterfactual density because the kink-induced compression of incomes toward k overcomes the notch-induced depression in the density, which becomes diffuse when lumpiness is high. In such a scenario, the fraction of unresponsive taxpayers in the KW model would not be well-defined; taken literally, the calculation of a^* would result in a negative value.

Together, these findings suggest that our estimation method is complementary to that of KW. The KW method is well-suited to settings where a subset of agents are unresponsive to a notch due to some frictions, and in such settings it provides a useful nonparametric quantification of those frictions in the form of the unresponsiveness share a^*. On the other hand, in settings where sparsity-based frictions are present, our estimation method can be used to quantify the elasticity and an alternative quantification of frictions in the form of the lumpiness parameter.

H Details of the South African tax system

The corporate income tax schedule for small business corporations and the personal income tax schedule are displayed in Figures A11a and A11b for selected years. In this appendix, we describe these tax systems and their rules in greater detail.

In South Africa, small business corporations (SBCs) face a progressive schedule of marginal tax rates that are lower than those applied to other firms. Figure A11a displays the schedule of marginal tax rates in 2018. Table A1 reports the full schedule of SBC tax rates in each year from 2010 to 2018. In addition to qualifying for lower tax rates, SBCs are also eligible for an accelerated depreciation allowance, and they are granted more generous deductible allowances for movable assets.

Businesses are eligible to register as an SBC if they meet each of the following requirements:[41]

- The business is a close corporation, co-operative, private company or personal liability company.[42]

[41] More information on these requirements can be found in an interpretation note provided by SARS at `https://www.sars.gov.za/wp-content/uploads/Legal/Notes/LAPD-IntR-IN-2018-08-Arc-08-IN9-Issue-6-Small-Business-Corporations.pdf`.

[42] A close corporation is a firm that was required to have 10 or less owners. After 2019, new companies could no longer incorporate as close corporations, but previously registered close corporations could maintain this form.

- All shareholders are natural persons (i.e, individuals and not companies or other legal structures) during the year of assessment.
- No shareholders may hold any shares or hold any interest in any other company, subject to certain exemptions. Some of these exemptions include listed companies, collective investment schemes and venture capital companies.
- The gross income of the company must not exceed R20 million for the year of assessment.
- The company may not be a personal service provider.[43]
- Investment income and income from rendering personal services may account for a maximum of 20 percent of all receipts, accruals and capital gains.[44]

SBCs account for approximately 26 to 31 percent of the total number of corporate tax filings between 2010 and 2018. Table A2 compares SBCs to other types of businesses in South Africa. It reports summary statistics for three groups of firms: non-SBCs, SBCs and size-matched non-SBCs, where the latter group consists of non-SBC businesses with revenues below the R20 million SBC eligibility threshold. Size-matched non-SBCs therefore comprise two types of firms: (i) firms who are eligible to register as an SBC but do not, either intentionally or because they are are unaware of the SBC program, and (ii) firms who are eligible to register as an SBC under the gross revenue requirement but who do not meet one of the other requirements listed above. We are unable to distinguish between these two types of firms in our data. While SBCs account for 38 percent of all companies, size-matched non-SBCs account for over half of all companies we observe. This discrepancy can be accounted for by recognizing that, since firms are not taxed when making losses and the number of loss-making firms greatly outnumbers the number of profit-making firms, many SBC-eligible firms do not register given that they make a loss and as such there is no incentive to SBC status. The size of the SBC sector is therefore a subset of the eligible SBC population, which would be closer in size to the total number of all small- and medium-sized enterprises (SMMEs), which stands at over 90 percent of all formally registered companies.[45]

The share of SBCs has risen over time. Figure A10a shows a clear upward trend since 2013, with SBCs accounting for over 31 percent of all tax filings in 2018. While large in number, the contribution of SBC tax revenue to total corporate income tax revenue, as shown in Figure A10b,

[43] Personal service providers refer to companies that have less than 3 employees and where more than 80 percent of the company income is derived from one client.

[44] Personal services refer to any company services performed personally by any person who holds an interest in that company when that company employs less than 3 employees. In this scenario, the tax authority deems the income being generated to be a function of the personal skill of that individual and not the company.

[45] This only takes into account formally registered firms. Given South Africa's large informal economy, the true number of SMMEs will be even larger.

is more modest, with SBCs accounting for 3 percent of overall corporate tax revenue on average during our sample period. This share has however increased since 2014, rising from around 2.5 percent to 4 percent. The jump in share of tax revenue from SBCs between 2013 and 2014 coincides with the year in which the gross income requirement for SBC eligibility was increased from R14 million to R20 million; this allowed a larger number of companies to register as SBCs and therefore increased the fraction of corporate tax revenue originating from SBCs.

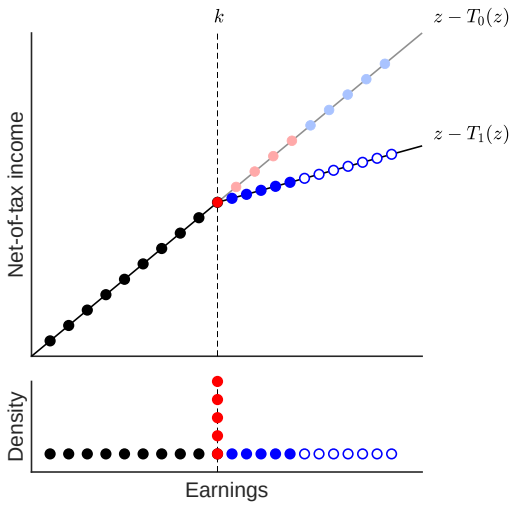
Personal income taxes represent the single largest source of tax revenue in South Africa. In 2017, the personal income tax base accounted for 34.2 percent of total tax revenue in the country. This figure is considerably higher than the OECD average of 23 percent, nearly twice as large as the average share for Africa (18 percent) and more than three times as large as the average share in Latin America (9.2 percent).

The personal income tax base comprises both wage-earners and the self-employed and extends both to personal earned income (wage and non-wage) and fringe benefits. Capital gains are taxed separately. Taxes are imposed at the individual level, i.e., spousal income does not affect individual tax liability. Deductions are permitted from taxable income; the most common deductions are from retirement contributions and expenses (such as travel expenses) incurred when producing income.

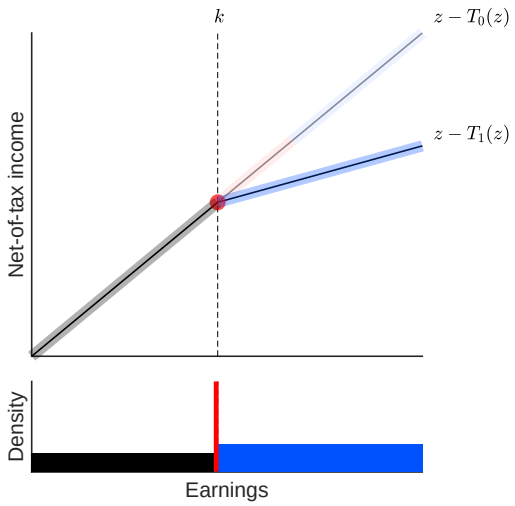
I Appendix Tables and Figures

Figure A1: Bunching patterns under a frictionless model

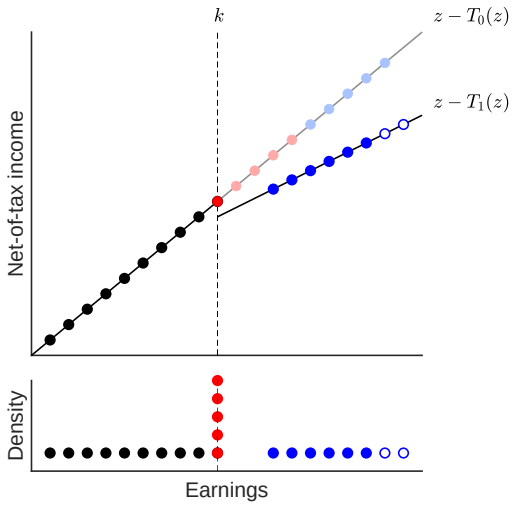
(a) Income choice around kink



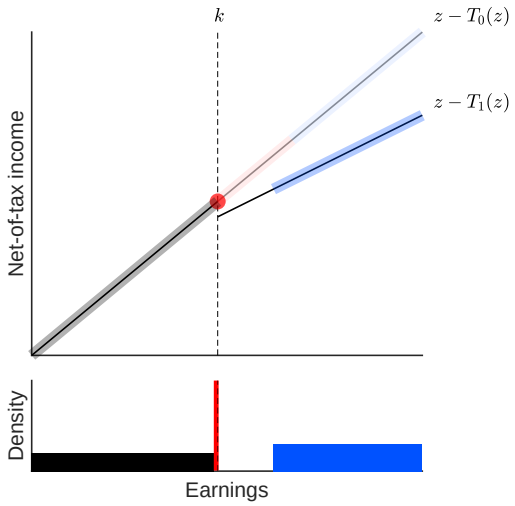
(b) Income density around kink



(c) Income choice around notch

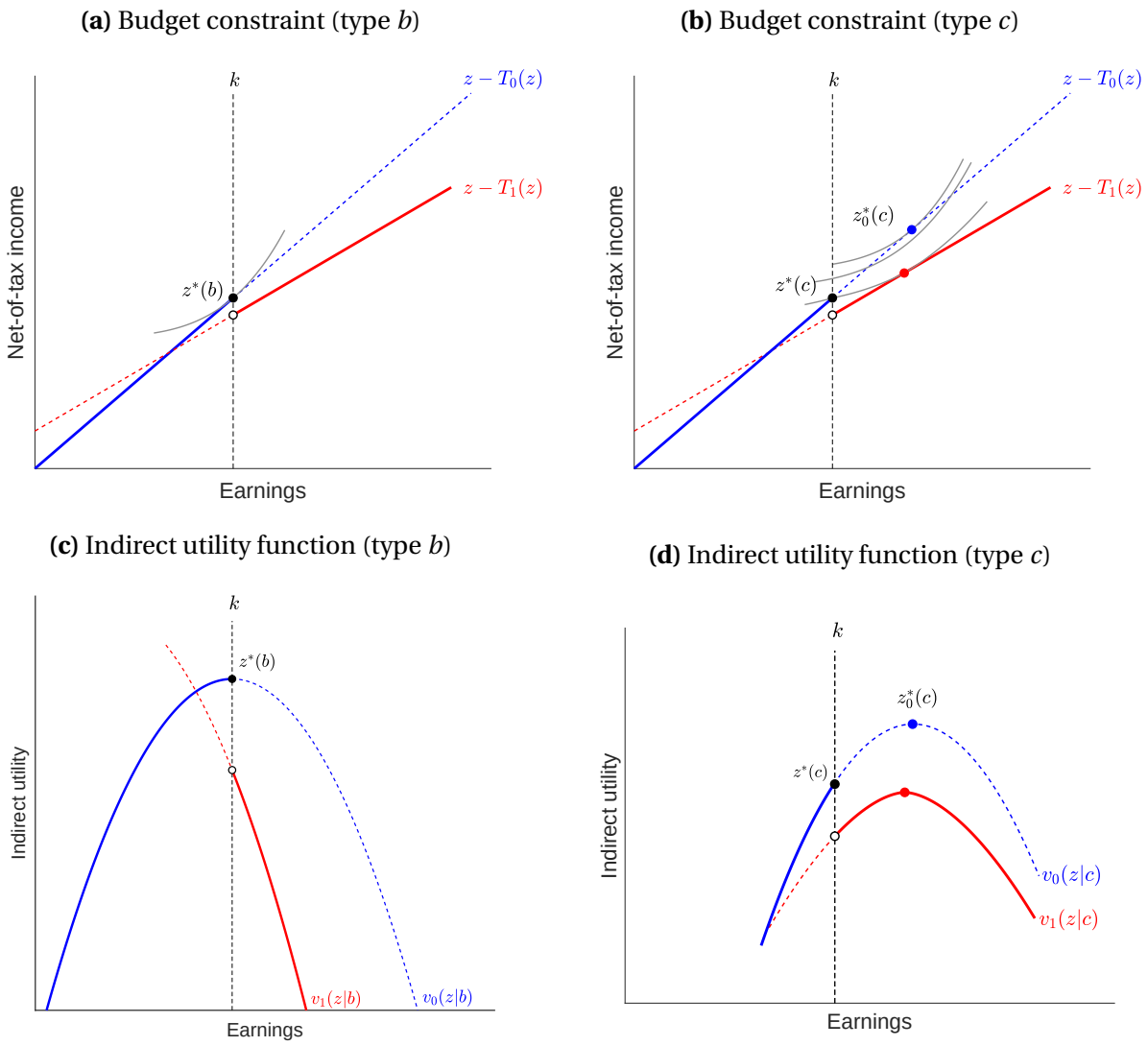


(d) Income density around notch



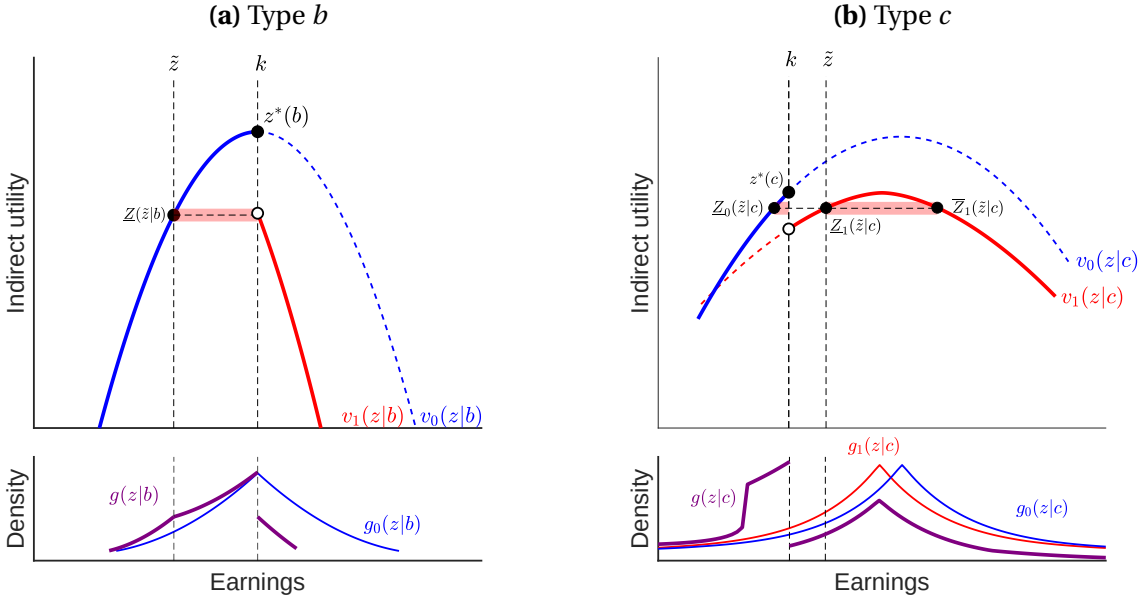
Panel (a) plots income choices for discrete types in the presence of a kink, for a uniform type distribution. Black points denote types who choose the same income under the linear tax $T_0(z)$ and the kinked tax schedule. Red points denote types who bunch at the bracket threshold k under the kinked tax schedule; their counterfactual income choices under $T_0(z)$ are plotted in light red for reference. Blue points denote types who choose incomes above the threshold under the kink. Hollow blue points denote agents whose counterfactual incomes under $T_0(z)$ lie outside the displayed range of incomes. The lower portion of Panel (a) displays the observed probability density function from these choices. Panel (b) translates to the case of continuous types, which exhibits an atom of mass at the threshold k and a jump in the density around that threshold, due to the compression of incomes in response to the higher marginal tax rate above the kink. Panels (c) and (d) are the same as Panels (a) and (b), but in the presence of a tax notch.

Figure A2: Utility from income choices around a tax notch



This figure is analogous to Figure 4, but in the presence of a notch, which produces a discontinuity in the indirect utility function (Panels (c) and (d)). In the case of type c, the notch produces a non-monotonic indirect utility function with two local maxima (Panel (d)).

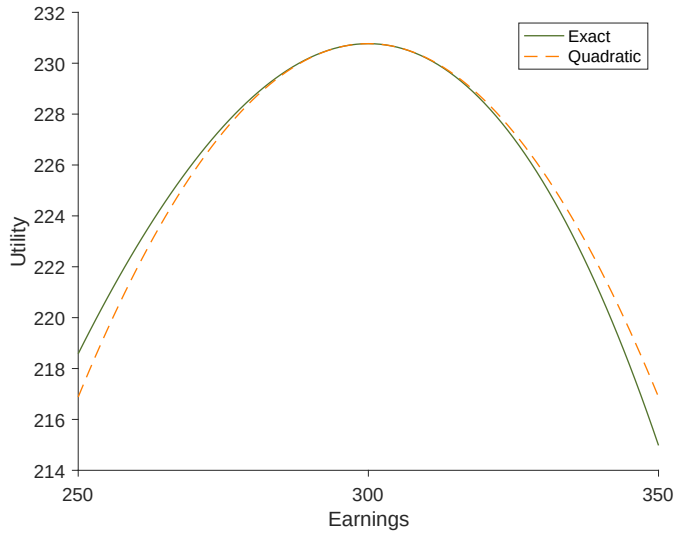
Figure A3: Type-conditional income density around a notch



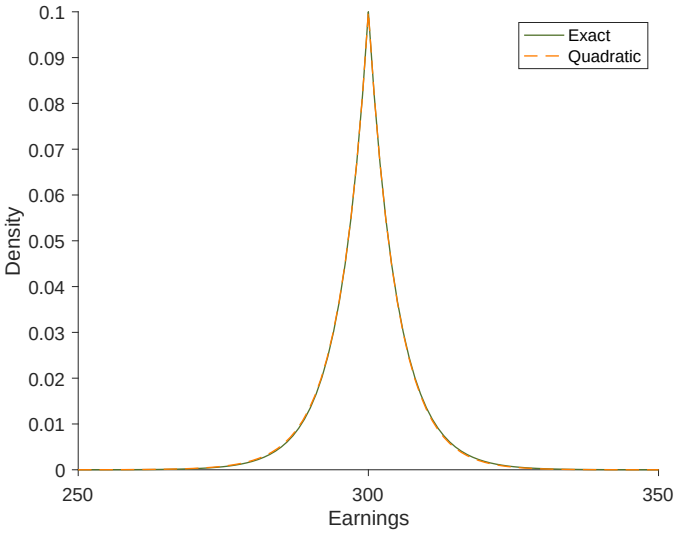
This figure is analogous to Figure 5, but in the presence of a notch. As shown in Panel (b), when the indirect utility function has multiple local maxima, the dominating income range may be a disjoint set, in which case the type-conditional density is multimodal.

Figure A4: Quadratic approximations of utility

(a) Indirect utility vs. quadratic approximation



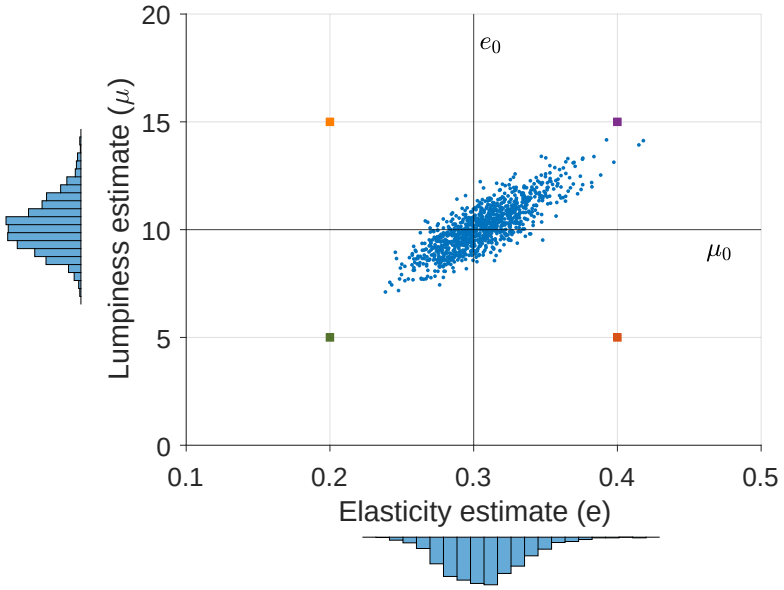
(b) Type-conditional income density vs. quadratic-utility approximation



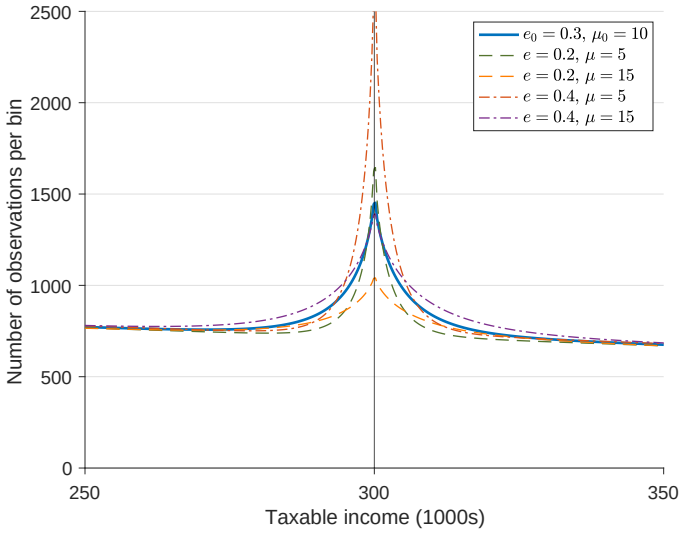
This figure illustrates the quantitative effect of imposing Assumption 1. Panel (a) plots exact indirect utility under a linear tax for the specification used in the simulations from Section 3, wherein taxpayers have constant elasticity of taxable income, and compares it to the quadratic approximation arising from the second-order Taylor approximation of that indirect utility around the taxpayer's target income. The approximation can be seen to be quite accurate across incomes near the target. Panel (b) plots the type-conditional income density under the linear tax and compares it to the density produced by the quadratic approximation. This approximation is extremely accurate, because the quadratic approximation in Panel (a) only diverges meaningfully from true utility at incomes far from the target, which are very unlikely to be chosen from the opportunity set.

Figure A5: Joint identification of elasticity and lumpiness estimates

(a) Joint distribution of $\hat{e}$ and $\hat{\mu}$ estimates

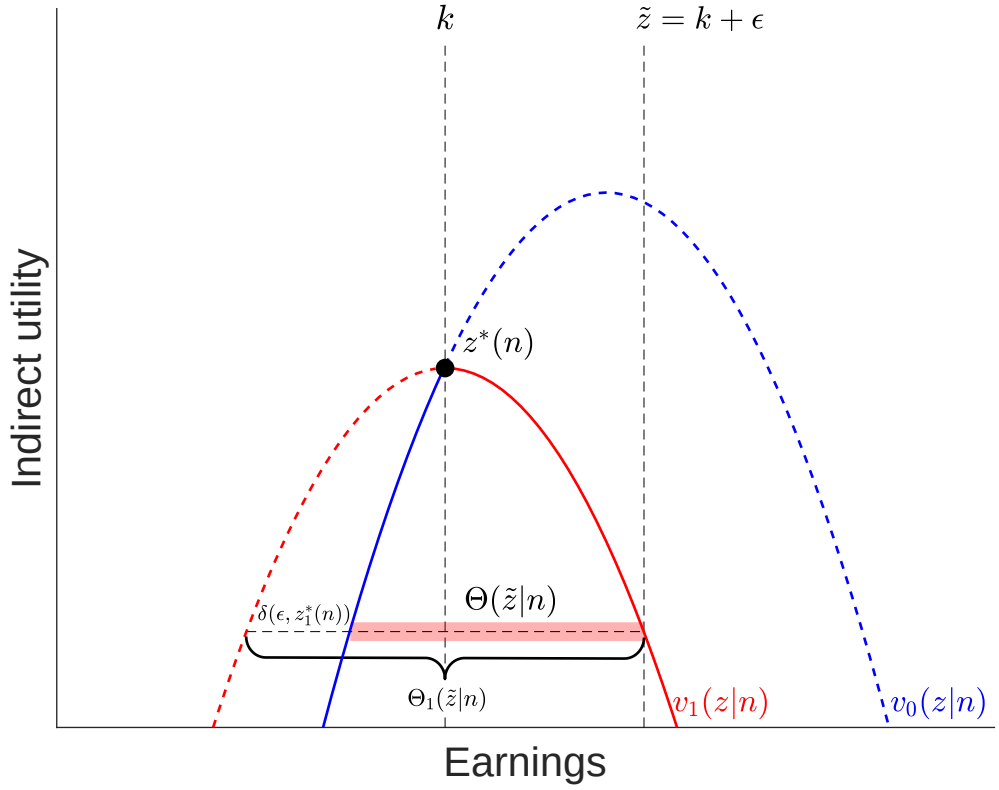


(b) Income densities for different combinations of e and μ



In Panel (a), each blue point plots the combination of parameter estimates $(\hat{e}, \hat{\mu})$ from one round of simulated data like that in Figure 9a. Marginal histograms of the estimates are plotted for each axis. Panel (b) plots the model-generated income density under the true parameters of the data-generating process, $e_0 = 0.3$ and $\mu_0 = \$10,000$, as well as under the four different combinations corresponding to the colored square points in Panel (a).

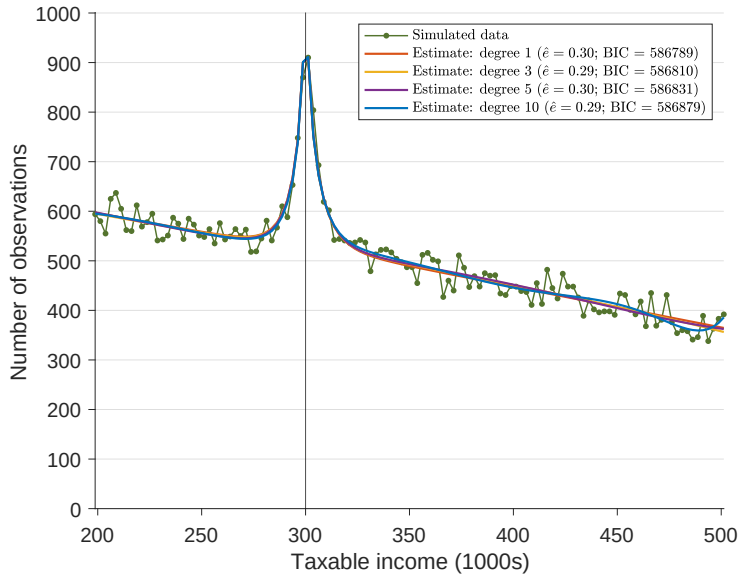
Figure A6: Illustration of dominated income regions with and without kink



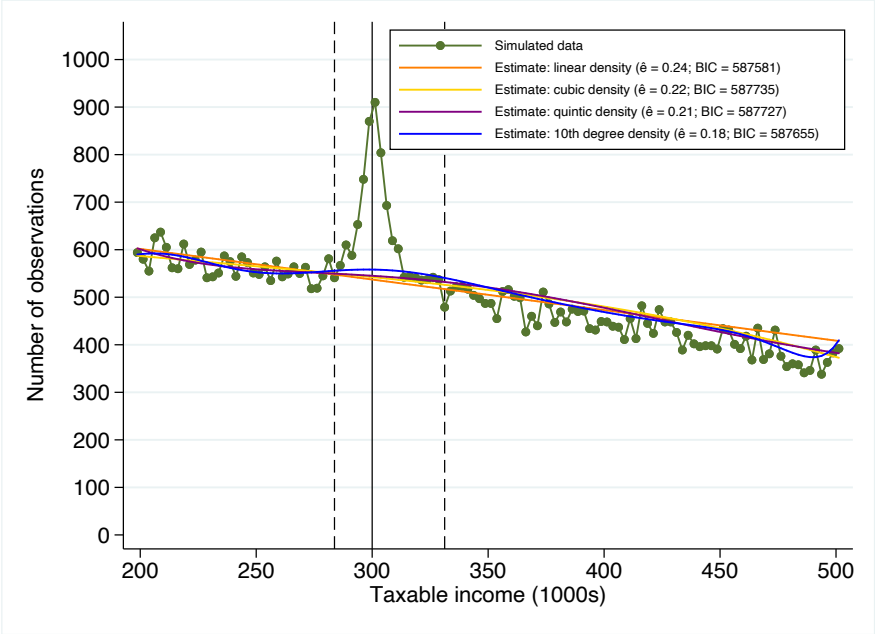
This figure illustrates definitions introduced in the identification proof in Appendix C. Θ represents the set of incomes that utility-dominate income opportunity $\tilde{z}$ in the presence of a tax kink at k. Θ_1 represents the set of incomes that dominate $\tilde{z}$ under the linear tax $T_1(z)$ which applies above the kink. δ represents the difference between these sets.

Figure A7: Estimated elasticities assuming different polynomial degrees

(a) Estimator with frictions



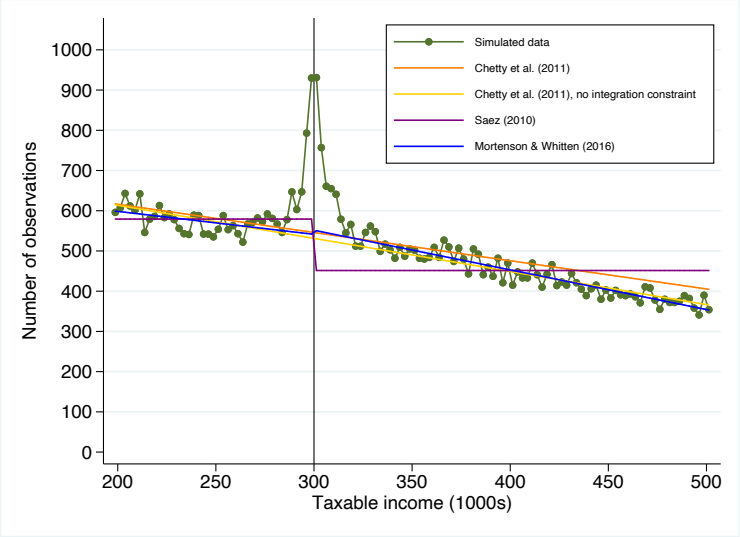
(b) Conventional estimator



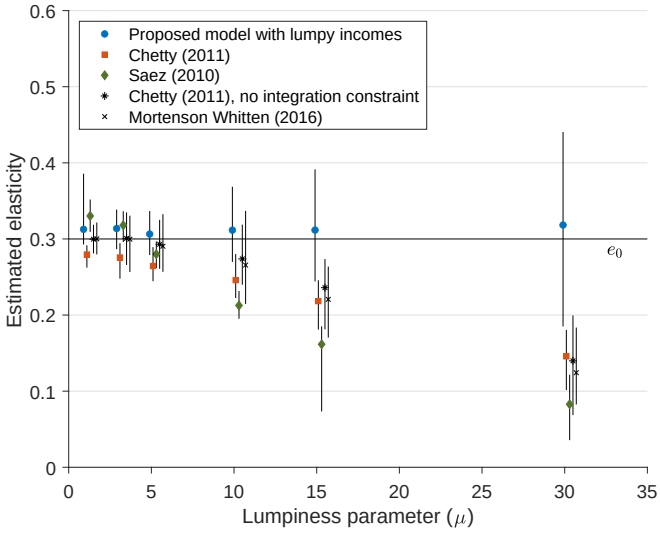
This figure reports the estimated elasticity for one round of simulated data, plotted in green, using both the conventional approach with frictionless income choice (Panel a) and our estimation method with lumpy income choice (Panel b), assuming different polynomial degrees for the counterfactual (or ability) density. The true ability density of the data-generating process is linear, with a true elasticity value of $e_0 = 0.3$.

Figure A8: Counterfactuals and elasticity estimates using various conventional approaches and our approach, for varying lumpiness parameters

(a) Alternative approaches to constructing counterfactuals



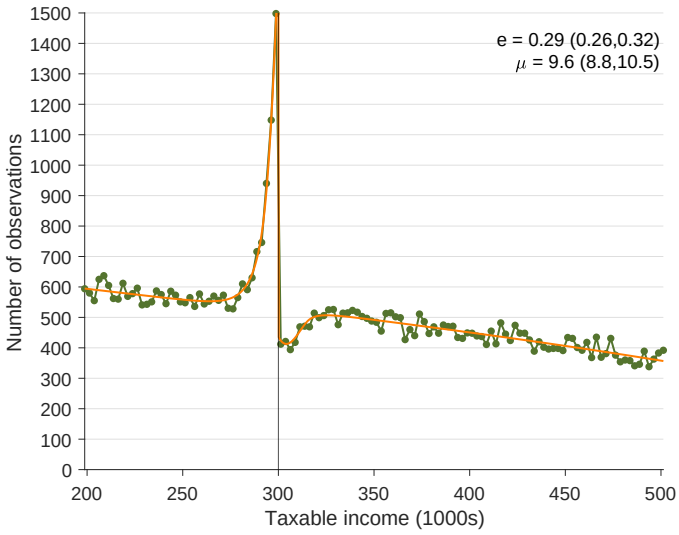
(b) Comparing the elasticity estimates



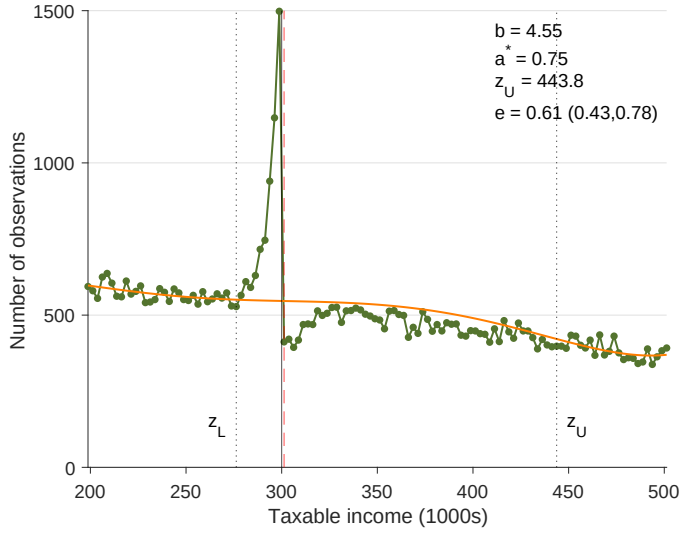
In Panel (a), we illustrate the counterfactuals produced under four different conventional bunching approaches to estimating elasticities for a simulated dataset where $\mu = 10$. In Panel (b), we simulate 100 rounds of data using a constant elasticity $e_0 = 0.3$ at each value of the lumpiness parameter μ_0 shown in the plots. We then estimate the elasticity $\hat{e}$ using our estimation approach and four conventional bunching estimators. The vertical lines indicate the 95 percent confidence intervals for the $\hat{e}$ estimates. For the conventional methods, we adapt the automated bunching window approach in Bosch, Dekker and Strohmaier (2020) in order to account for each method's approach to constructing a counterfactual distribution.

Figure A9: Elasticity estimation based on Kleven and Waseem (2013)

(a) Simulated data with notch, estimated using our proposed model with frictions



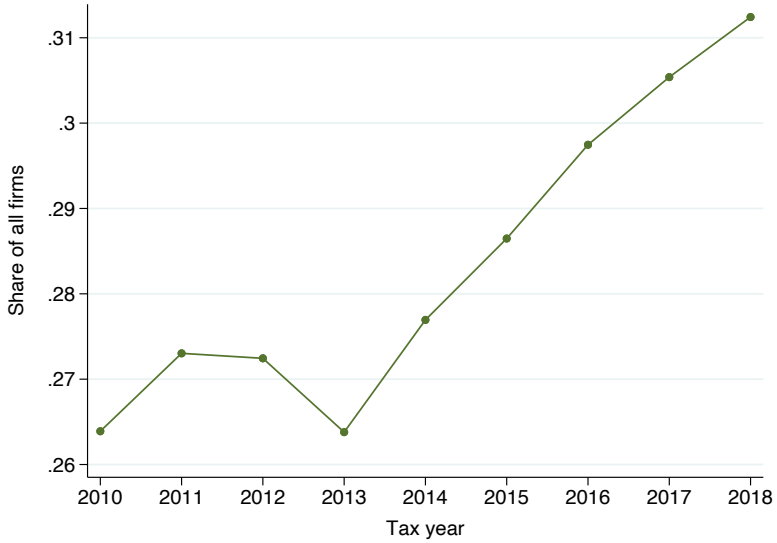
(b) Estimation based on Kleven and Waseem (2013)



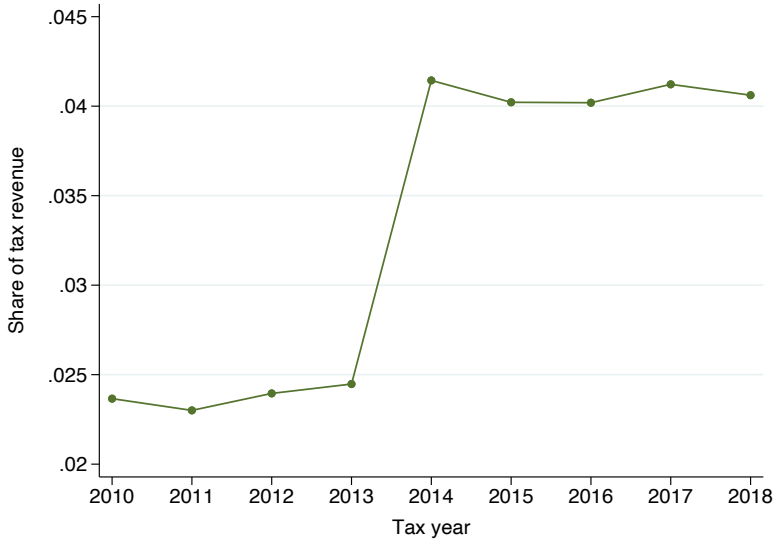
Both panels plot the same round of simulated data with sparsity-based frictions, using an elasticity of $e_0 = 0.3$ and a lumpiness parameter of $\mu_0 = 10$, and a notch value of \$1000; other tax parameters are the same as in Figure 9. Panel (a) applies our maximum likelihood estimator with frictions. Panel (b) applies the Kleven and Waseem (2013) notch-based bunching estimator. The vertical dashed line just above the bracket threshold indicates the upper bound of the "dominated income region." b is the average excess mass between between the visually specified lower bound of the bunching window z_L and the threshold k, in proportion to the average estimated counterfactual frequency (shown in orange) in the dominated income region. a^* is the empirical frequency in the dominated region, as a share of the counterfactual density. z^U represents the upper bound of the bunching region, which is computed so that the missing mass equals the excess bunching mass.

Figure A10: Small Business Corporation (SBC) prevalence and contribution to corporate income tax revenue

(a) SBCs as a share of all firms



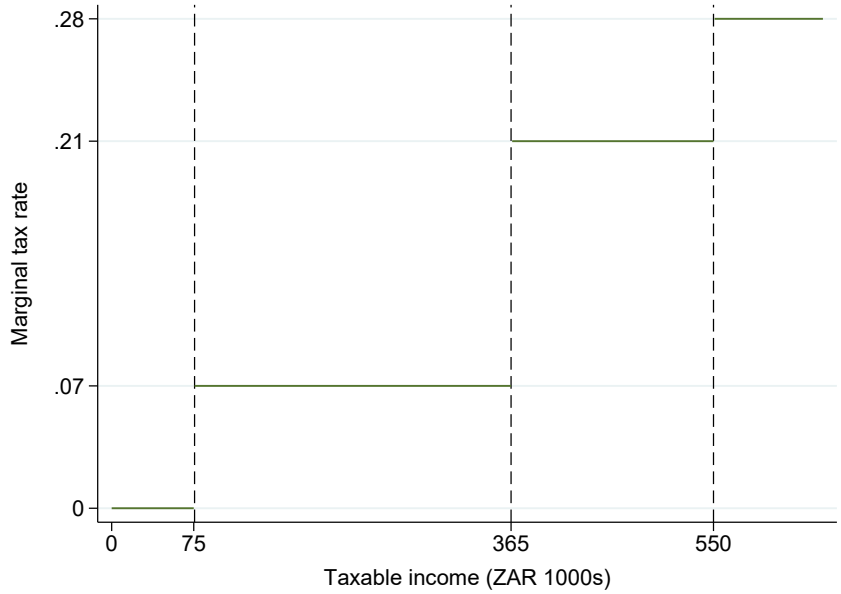
(b) SBCs contribution to total corporate income tax revenue



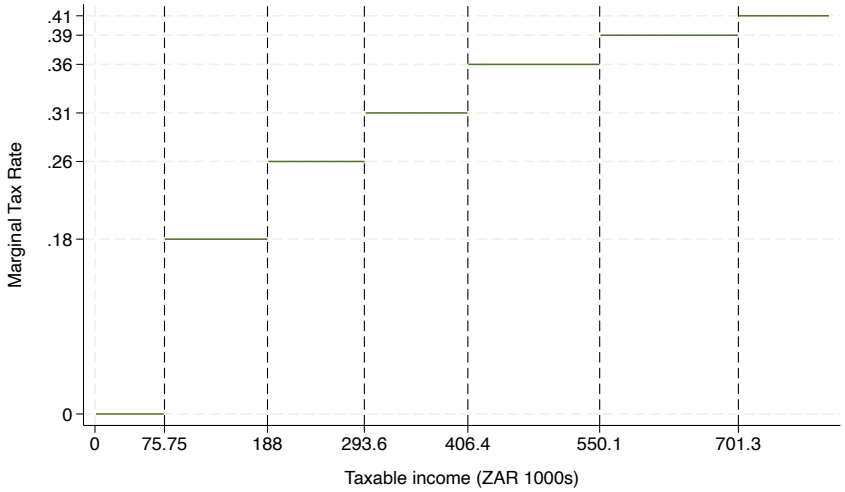
Panel (a) shows the share of SBC tax filings relative to all corporate tax filings between tax years 2010 and 2018. Panel (b) shows the share of corporate tax revenue contributed by SBC's as a percentage of total tax revenue between tax years 2010 and 2018. In 2014, the income ceiling for SBCs was raised from R14 million to R20 million, generating a substantial increase in their contribution to total revenues. To calculate tax revenue, we sum up the tax liability of firms. We do not observe whether a payment was made and as a result, the figure should be viewed as indicative of taxes owed.

Figure A11: Marginal Tax Rate Schedules in South Africa

(a) Tax Schedule for Small Business Corporations in 2017

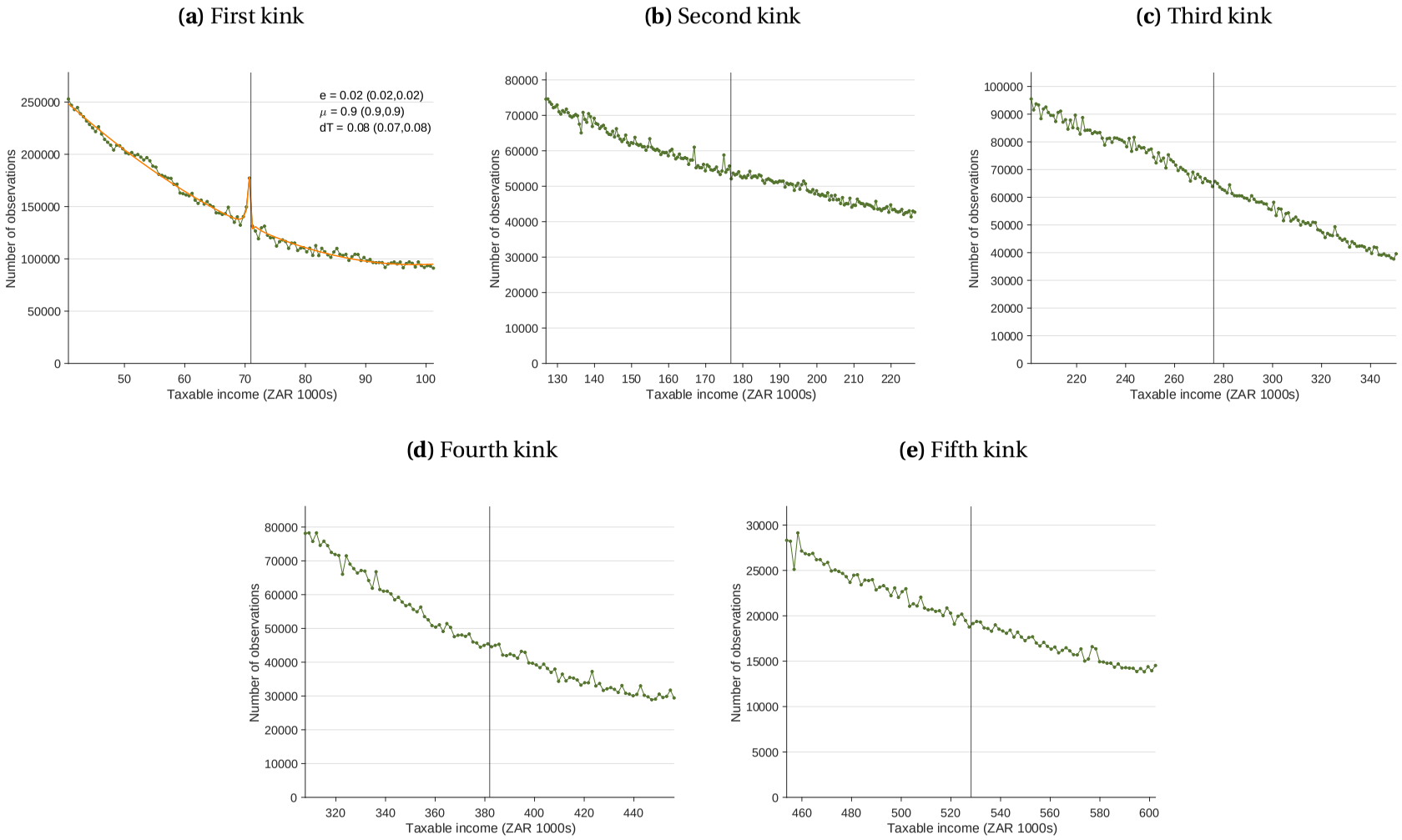


(b) Personal Income Tax schedule in 2017



Panel (a) shows the marginal tax rate schedule for small business corporations in South Africa in 2017. The horizontal axis is measured in South African rand (ZAR), and vertical dashed lines indicate bracket thresholds where marginal tax rates change. Panel (b) shows the personal income taxation marginal tax rate schedule in 2017.

Figure A12: Model estimation: personal income taxes (wage earners)



Green points plot the empirical histogram of individuals who earn a wage from an employer with different earnings in the data and orange lines plots the predicted density generated by the maximum likelihood estimation of the uniform sparsity model parameters e (elasticity of taxable income), μ (average distance between income opportunities in ZAR 1000s), and dT (the estimated "as-if" discrete change in tax liability at the bracket threshold, in ZAR 1000s). Numbers in parentheses indicate the 95 percent confidence interval on parameter estimates generated by the MLE method.

Table A1: Small Business Corporation Tax Schedule, 2010–2018

Tax year	Taxable income	Marginal tax rate
2010	0 - 54,200	0%
	54,200 - 300,000	10%
	Above 300,000	28%
2011	0 - 57,000	0%
	57,000 - 300,000	10%
	Above 300,000	28%
2012	0 - 59,570	0%
	59,570 - 300,000	10%
	Above 300,000	28%
2013	0 - 63,556	0%
	63,556 - 350,000	7%
	Above 350,000	28%
2014	0 - 67,111	0%
	67,111 - 365,000	7%
	365,001 - 550,000	21%
	Above 550,000	28%
2015	0 - 70,700	0%
	70,700 - 365,000	7%
	365,001 - 550,000	21%
	Above 550,000	28%
2016	0 - 73,650	0%
	73,651 - 365,000	7%
	365,001 - 550,000	21%
	Above 550,000	28%
2017	0 - 75,000	0%
	75,001 - 365,000	7%
	365,001 - 550,000	21%
	Above 550,000	28%
2018	0 - 75,750	0%
	75,751 - 365,000	7%
	365,001 - 550,000	21%
	Above 550,000	28%

This table indicates the small business corporation (SBC) graduated income tax system for the tax years 2010–2018. Tax years run from April 1 to March 31.

Table A2: Summary statistics for businesses filing corporate income tax returns, 2014–2018

Company type	Non-SBC	Size Matched Non-SBC	SBC
Turnover (in R'000)	121,249.1	3,028.22	2,628.6
	(521,791.0)	(4,275.89)	(3,531.4)
Expenses (in R'000)	33,713.74	1,684.13	1,244.78
	(281,924.3)	(2,632.84)	(1,785.2)
Assets (in R'000)	73,928.56	3,422.88	1,264.86
	(658,585.4)	(16,459.2)	(2,344.89)
Liabilities (in R'000)	51,659.0	2,231.0	673.0
	(510,496.1)	(13,190.15)	(1,708.29)
Inventory (in R'000)	10,782.82	284.19	173.72
	(69,005.84)	(2,017.28)	(651.83)
Cash (in R'000)	7,185.93	312.86	199.56
	(47,614.75)	(1,606.54)	(630.77)
Net profit (in R'000)	5,280.92	92.31	125.68
	(31,990.6)	(1,066.98)	(502.82)
Number of employees	90.75	4.97	3.93
	(579.44)	(19.11)	(11.69)
Number of salaried directors	2.26	1.47	1.32
	(5.64)	(0.83)	(0.62)
Taxable income (in R'000)	-462.12	-346.43	6.39
	(150,598.8)	(3,541.95)	(753.68)
Tax liability (in R'000)	1,510.64	49.41	30.56
	(6,659.3)	(176.17)	(119.75)
% of firms with a salaried director	35.18%	13.70%	17.34%
% of firms with a tax practitioner	73.09%	71.12%	64.16%
Number of unique tax returns	137,872	653,755	457,198
Share of tax revenue	81.82%	12.69%	5.49%
Number of unique companies	41,289	238,830	172,440
Share of companies	9.12%	52.77%	38.10%

This table reports summary statistics for corporate income tax returns in South Africa between 2014 and 2018 for 3 groups of firms: "Non-SBCs," "Size Matched Non-SBCs" and "SBCs." "Size Matched Non-SBCs" represent "Non-SBCs" with revenues below R20 million, the SBC eligibility threshold. "Size Matched Non-SBCs" and "Non-SBCs" are mutually exclusive categories. Standard deviations are shown in parentheses.

Table A3: Estimated bunching parameters: personal income taxes

(a) Parameter estimates from model with frictions

	Elasticity of taxable income (e)		
	First kink	Second kink	Third kink
Wage earners	0.02 (0.02, 0.02)	—	—
Self-employed	0.53 (0.53, 0.54)	0.02 (0.02, 0.03)	0.03 (0.02, 0.03)
	Lumpiness parameter (μ), in ZAR 1000s		
	First kink	Second kink	Third kink
Wage earners	0.9 (0.9, 0.9)	—	—
Self-employed	4.4 (4.3, 4.5)	1.8 (1.6, 1.9)	4.5 (4.1, 5.0)
	As-if notch value (dT), in ZAR 1000s		
	First kink	Second kink	Third kink
Wage earners	0.08 (0.07, 0.08)	—	—
Self-employed	0.31 (0.31, 0.31)	0.06 (0.04, 0.07)	0.02 (-0.01, 0.05)

(b) Parameter estimates from conventional bunching estimator

	Elasticity of taxable income (e)		
	First kink	Second kink	Third kink
Wage earners	0.03 (0.03, 0.04)	—	—
Self-employed	0.44 (0.43, 0.45)	0.03 (0.03, 0.03)	0.02 (0.02, 0.02)

Panel (a) reports our maximum likelihood estimates of the elasticity of taxable income (e), the average distance between income adjustment opportunities (μ) and the revealed preference ("as-if") value of the change in tax liability at each bracket threshold. The values of μ and dT are measured in ZAR 1000s. Results are reported separately for the aggregate population, and for the subset of individuals who are wage earners and those who are self-employed. Panel (b) reports the estimated elasticity (e) from the conventional bunching estimator, using the method based on Chetty et al. (2011) and described in Appendix D.

Table A4: Comparability of aggregate estimates to estimated aggregated up from subsamples

	MAPE	Aggregate elasticity (e)	Weighted elasticity across subsamples (e)	Aggregate lumpiness parameter (μ)	Weighted lumpiness parameter across subsamples (μ)
Lower kink	0.20%	1.75 (1.72, 1.79)	1.79	5.9 (5.7, 6.1)	6.1
Middle kink	0.27%	0.27 (0.24, 0.31)	0.28	11.3 (9.1, 13.5)	14.2
Upper kink	0.36%	0.23 (0.19, 0.28)	0.26	6.6 (4.7, 8.5)	10.6

The table reports the comparability of our model parameters estimated on the aggregate (or full) sample of corporate income tax returns to the model parameters estimated on the subset of firms who do and do not use paid tax practitioners to prepare their tax returns which we then aggregate together. In the first column we report the Mean Average Percentage Error (MAPE) which represents the percentage difference between the full sample model fit and the model fit obtained by aggregating the two CIT sub-samples. Column two reports the aggregate sample elasticity (e) estimates. Column three reports the elasticities generated by a weighted average of the elasticites produced on each subsample, weighted by the share of each subsample in the full sample. Column four reports the average distance between income adjustment opportunities (μ) estimated on the full sample. Lastly, column five reports the weighted μ estimated on each subsample, weighted by the share of each subsample in the full sample.