Diffuse Bunching with Frictions: Theory and Estimation*

Santosh Anagol Allan Davids
Benjamin B. Lockwood Tarun Ramadorai†

October 13, 2022

Abstract

We incorporate a model of frictions into the bunching-based elasticity estimator to rationalize diffuse bunching around kinks and mass above notches in empirical distributions. Model agents draw a sparse set of opportunities from a Poisson process, approximating a broad class of frictions including search costs, inattention, and lumpy adjustment; the predicted density depends on the standard structural elasticity and a money-metric "lumpiness parameter." We estimate the model using administrative tax data on South African small-businesses, recovering moderate elasticities of taxable income between 0.2 and 0.3 at higher incomes, and larger elasticities at low incomes. Firms appear to treat the bottom kink as a notch, and firms with paid tax practitioners exhibit sharper bunching, driven primarily by lower frictions rather than a higher elasticity.

*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, the Penn Wharton Budget Model, and the Wharton Dean's Research Fund 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, Henrik Kleven, Dylan Moore, Jacob Mortenson, Alex Rees-Jones, Joel Slemrod, Jakob Søgaard, David Thesmar, Andrew Whitten, Eric Zwick, and conference and seminar participants at Economic Research South Africa (ERSA), the European Bank for Reconstruction and Development, Imperial College, LAGV 2021, LMU Munich, CREST, the Toulouse School of Economics, the Wharton Applied Economics Workshop, and the South African Revenue Services for helpful comments and to 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 and the revenue-maximizing tax rate. As such, it is the subject of a large body of empirical research (Saez, Slemrod and Giertz, 2012). One prominent 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 kinks—points where the marginal tax rate steps up at a tax bracket threshold—using a formula derived from a simple model of taxpayer optimization. Kleven and Waseem (2013) extend this approach to handle notches, where the tax level (rather than the marginal rate) steps up.

This strategy has proven influential in part because of its light data requirements: bunching estimation can be performed using only cross-sectional, anonymized counts of tax returns at each income level. Indeed, this influence can be seen in the wide application of bunching estimators beyond the setting of income taxation. They have been used to quantify behavioral responses in the 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). Bunching-based estimation has also been used to uncover evidence of reference dependence in settings such as tax refunds (Rees-Jones, 2018), marathon times (Allen et al., 2017), and house selling (Andersen et al., 2022).

The standard bunching estimator formula takes as its input the excess mass in an income (or other) density around a kink or notch relative to the "counterfactual density" that would obtain if the tax schedule were linear. The formula is derived using a frictionless model that predicts an atom of excess mass precisely at the kink, in which case the counterfactual density is revealed by the income density immediately beside that atom of excess mass.

In contrast to the frictionless model, the bunching observed in empirical distributions is typically diffuse. This poses an important estimation challenge in this context because it obscures the distinction between bunching mass induced by the tax schedule and fluctuations in the underlying counterfactual density. This issue has been discussed and studied extensively in the bunching literature. Chetty et al. (2011) propose a frequently applied solution to this problem, measuring 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. The elasticity is then computed from this diffuse bunching mass, effectively proceeding as if the diffuse mass were counterfactually located precisely at the kink.[1]

[1] Mortenson and Whitten (2020) apply this estimation method to the setting of U.S. taxpayers bunching in re-

In the case of notches in the tax schedule, Kleven and Waseem (2013) propose a non-parametric approach to estimate frictions using the mass located in the strictly dominated region. However, neither of these approaches provides a formal model of the frictions that produce diffusion.

In this paper, we extend these approaches by incorporating a parsimonious but general model of frictions into the bunching estimator. This allows us to improve estimates of the counterfactual density, which we show can be confounded by the presence of frictions under the standard approach, leading to potentially substantial underestimates of the elasticity. Moreover, by treating diffusion as an informative feature of the data, rather than an artifact to be eliminated, we are able to draw novel economic insights about the strength of frictions and the money-metric size of unobserved optimization costs.

We consider the class of models with "sparsity-based frictions," in which agents choose between discrete opportunities drawn from around their frictionless target. This structure accommodates a broad range of frictions that have been discussed as likely contributors to observed diffusion in bunching, 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), which can be viewed as different microfoundations for the opportunity generation process.[2] We show that these models share a common structure and are well approximated by a limiting case—in a sense formalized in our Proposition 1—of "uniform sparsity," in which opportunities are drawn from a Poisson process.

Our uniform sparsity model distills the many details of the opportunity generation process, such as the number of opportunity draws and the density from which they are drawn, into a single sufficient statistic—the "lumpiness parameter"—which quantifies the expected difference between adjacent opportunities around an agent's target. This model avoids the need to estimate (or visually specify) bounds for the excluded bunching window, so the total number of model parameters is reduced relative to the conventional approach. Despite this parsimony, the interaction between optimizing agents and sparse opportunity sets produces rich predictions about the shape of bunching around kinks and notches, which match key features of empirical bunching patterns.

In the presence of a kink, agents in the model who target the kink ("would-be bunchers") select their closest available opportunity, producing approximately symmetric diffuse bunching

sponse to the Earned-Income Tax Credit. Bosch, Dekker and Strohmaier (2020) and Dekker and Schweikert (2021) study the optimal selection of the excluded bunching window. See Kleven (2016) for a review of the bunching estimation literature.

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

around the bracket threshold. Around a notch, would-be bunchers who draw an opportunity just below the bracket threshold will choose that opportunity. However, those with an opportunity just above the threshold will often prefer a more distant opportunity in order to avoid incurring the penalty from the notch. This produces asymmetry in the bunching mass, with a depression in the observed density above the notch. We analytically characterize the predicted income density as a function of the elasticity, the lumpiness parameter, and a parameterized smooth underlying ability density, which allows for maximum likelihood estimation.

By incorporating a positive model of frictions, this estimation strategy advances bunching estimation methods in two respects. First, we improve the estimation of the elasticity itself. Using simulations produced with sparsity-based frictions, we confirm that our method recovers the true elasticity of the data-generating process with accurate confidence intervals. In the same simulations, the standard kink-based bunching estimator (Saez, 2010; Chetty et al., 2011) underestimates the true elasticity by over 50 percent in some specifications, and its confidence intervals can be overly precise: in some simulations, the 95 percent confidence intervals contain the true elasticity less than 10 percent of the time. This mismeasurement arises from the difficulty of estimating the counterfactual density using an excluded bunching window; in the presence of sparsity-based frictions, some excess mass spills beyond the specified bunching window, distorting upward the estimated counterfactual density in the region around the threshold and leading to an underestimate of the bunching mass and thus the elasticity.[3]

We perform a similar exercise with the notch-based estimation method of Kleven and Waseem (2013), which accounts for the presence of empirical mass in the dominated income range above a notch by assuming that a share of agents are unresponsive to the notch due to frictions, and scaling up observed bunching to correct for that unresponsive share. When that estimator is applied to an income distribution produced by sparsity-based frictions, we find that the rescaling procedure overestimates the structural elasticity. These results highlight the complementary nature of our method, which is based on a different model of frictions. The choice between our alternative notch-based elasticity estimator and the Kleven and Waseem (2013) approach can be informed by the models' different qualitative predictions about the shape of the induced bunching around a notch.

Second, our estimation procedure gleans additional information about economic behavior from the pattern of bunching around kinks and notches that is neglected when diffuse excess mass is captured as a scalar in the conventional method. Here, we are motivated by our empirical application, which involves understanding the bunching behavior of the taxable income reported by small businesses in South Africa around three prominent tax kinks. The histograms

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

of taxable income around each kink are displayed in Figure 1, where diffuse bunching is evident at each bracket threshold.

The new estimation strategy that we develop in this paper uncovers two novel economic insights about taxpayer behavior when applied to the South African administrative data. First, whereas the bunching patterns around the middle and upper kinks in Figure 1 look similar to those typically observed around tax kinks, the pattern around the lower kink resembles the shape usually observed around a tax *notch* (see, e.g., Kleven and Waseem, 2013) with excess mass to the left of the tax bracket threshold and missing mass to the right. Such behavior could arise from an unobserved discrete cost—whether real or behavioral—of having earnings above the lower kink, raising a natural question: how big is this "as-if" money-metric notch value? Standard bunching-at-a-notch estimation methods are not designed to answer this question, because they require specifying the notch value (and thus the dominated income region) as a prerequisite for estimating the elasticity. In contrast, we can use our model to estimate the "as-if" notch value, which is identified by the asymmetry in the diffusion of bunching mass around the threshold. This asymmetry is discarded when the bunching pattern is reduced to a single number under the conventional approach.

The second economic insight relates to heterogeneity in bunching responses across different types of taxpayers. In the South African tax data, there are notable visual differences in bunching patterns among firms that employ a paid tax preparer and those that do not: bunching in the former group appears more tightly concentrated around the kink. Despite these observable differences, the conventional bunching estimation approach does not detect differences behavior between the two groups, producing elasticity estimates that are statistically indistinguishable from each other. In contrast, our estimation method uncovers material differences between the elasticities of these two groups of firms. The greater diffusion in bunching mass among firms without paid tax preparers produces a higher lumpiness parameter for this group, consistent with a coarser degree of income targeting or less attention to income optimization.

Although our application focuses on bunching in response to the income tax, this estimation approach can readily be applied in other settings with kinked or notched budget sets. For example, many bunching applications involve thresholds in other tax instruments which combine statutory changes in marginal or discrete tax incentives with unknown—but potentially important—changes in behavioral frictions or compliance costs. Value-added-tax (VAT) exemption thresholds (as studied in Velayudhan, 2018; Liu and Lockwood, 2015) typically have both a known increase in the tax rate and an unknown change in compliance costs.[4] Estimat-

[4]The method can also be applied in settings where there is a known statutory notch value (as in Kleven and Waseem, 2013, where the average tax rate increases at certain thresholds); the analyst can estimate the notch value

ing the elasticity with respect to the VAT rate using standard bunching methods requires the researcher to assume a compliance cost, and then to use the residual bunching to estimate the elasticity. In our method, both the elasticity and the revealed-preference compliance cost can be estimated based on observed bunching behavior.[5]

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 results on the estimation of the tax notch value 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. On the question of confusing average and marginal tax rates, 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 literature measuring behavioral responses to taxation in developing economies. These include 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 Boonzaaier et al. (2019) in our setting of South Africa.

The rest of the paper proceeds as follows. In Section 2, we present our model of frictions, we characterize the resulting bunching density, and describe our proposed maximum likelihood estimation method. In Section 3, we compare the performance of our estimation method with the conventional bunching estimators using simulated data with known underlying parameters. In Section 4, we describe our empirical application to small businesses in South Africa and present our estimation results. Section 5 concludes.

using our method and then compare it to the known notch value. A comparison of the estimated and statutory notch values could then be used either as a specification test, or as an indicator of whether there is some other friction at the threshold.

[5]This method also extends to settings where the researcher wishes to estimate a notch value associated with some non-monetary choice behavior. For example, Allen et al. (2017) find bunching behavior among marathon runners at round-number time increments (e.g., 4 hours). In principle, our method could be used to estimate the notch value, measured in terms of minutes of marathon time, associated with achieving a given time threshold.

2 Model

2.1 Baseline bunching model with frictionless choice

Our starting point is the canonical "bunching estimator" presented in Saez (2010). Although this model naturally extends to many non-tax settings with kinked or notched budget sets, for concreteness we will use the language of income choice and income taxes for this exposition. In the standard setup, taxpayers have the following utility function:

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

When taxpayers are individual workers, c and z are interpreted as consumption and pre-tax earnings from labor effort, which are observable to the government. Taxpayers are heterogeneous in their income-earning ability, which is indexed by $n \in (0,\infty)$ and has distribution $F(n)$ and density $f(n)$. An n-type taxpayer chooses the level of z that maximizes utility in equation (1), subject to their budget constraint:

$$c = z - T(z), \tag{2}$$

where $T(z) = a + tz$ is a (locally) linear income tax.[6] Type n's constrained-optimal level of income $z^*(n)$ satisfies the following first-order condition:

$$z^*(n) = n(1-t)^e. \tag{3}$$

Thus n can also be interpreted as an individual's preferred income under a hypothetical laissez-faire tax system with $t = 0$. Equation (3) satisfies $\frac{d\ln z^*(n)}{d\ln(1-t)} = e$, so that e can be interpreted as the elasticity of taxable income with respect to the marginal net-of-tax rate $1-t$.

In the bunching literature, this model is used to estimate the elasticity e based on the behavior of the empirical income density, denoted $h(z)$, around a threshold between two different linear income tax segments. We employ the following notation to describe a piecewise-linear tax function $T(z)$ around a bracket threshold k:

$$T(z) := \begin{cases} T_0(z) = a_0 + t_0 z & \text{if } z \le k \\ T_1(z) = a_1 + t_1 z & \text{if } z > k \end{cases} \tag{4}$$

The identifying assumption underlying this strategy is that the density of types $f(n)$ is smooth—

[6] We do not model the choice of *whether* to file a tax return (i.e., an extensive margin response). See Pollinger (2021) for a bunching model that attempts to estimate both intensive and extensive margins combining bunching and regression kink design models.

in a sense that will be made precise—across the range of types with earnings near k, implying that observed deviations away from a smooth income density in the vicinity of k can be attributed to the change in tax rates at the bracket threshold. In the case of a progressive "kink" where the marginal tax rate increases at the bracket threshold ($t_0 < t_1$) but the tax level is continuous ($T_0(k) = T_1(k)$), this model predicts excess mass in the income density at k.

Figure 2 illustrates income choices induced by a progressive tax kink for four selected types of taxpayers, as well as their counterfactual income choices under the linear tax $T_0(z)$. Employing equation (3), taxpayers with n between $k/(1-t_0)^e$ and $k/(1-t_1)^e$—the "marginal non-buncher" and "marginal buncher," respectively, depicted by types b and c in Figure 2—all choose to earn exactly k. The marginal buncher's income choice under the kinked schedule is the same as it would be if the linear tax $T_1(z)$ applied everywhere, and thus the income change for the marginal bunching type c (from $z_0^*(c)$ to $z_1^*(c)$ in Figure 2c) in response to the difference in tax rates (t_1 vs. t_0) identifies the income elasticity e.

Figure 3 illustrates the observed income density around a tax bracket threshold in this model. In Panel (a), red points denote income choices for discrete taxpayer types between the marginal non-buncher and the marginal buncher; the lower portion of the panel illustrates the induced density function, with bunching types stacking up at the bracket threshold. Types above the marginal buncher reduce their incomes to a new interior optimum, resulting in compression of income choices to the right of k relative to the counterfactual choices under the linear tax $T_0(z)$. Panel (b) illustrates this behavior under a continuous type density, which exhibits an atom of mass at the bracket threshold and a discontinuity in the density around the threshold due to compression.[7] The empirical strategy of Saez (2010) amounts to estimating the "bunching mass" around the bracket threshold and using it to infer the marginal buncher's counterfactual income z_0^*. The details of this method are described in detail in Appendix A.2, where equation (32) reports the key formula for this inference.

Panels (c) and (d) of Figure 3 illustrate the predicted income density around a *notch*, where the tax level rises discontinuously at the bracket threshold.[8] As described in Kleven and Waseem (2013), in the presence of a notch, the threshold income level k strictly dominates incomes immediately above it, which require greater effort but produce lower net income. This creates a "hole" in the density immediately to the right of the threshold.

In contrast to the model-predicted densities in Figure 3, empirical income densities around bracket thresholds feature bunching that is diffuse and, in the case of notches, they exhibit pos-

[7] The density to the right of k also shifts leftward due to the kink. In the case of a uniform type density (illustrated here) this shift does not affect the density; however, if the type density is decreasing in the vicinity of the bracket threshold, then the upward discontinuity in the income density at k may be dampened or reversed.

[8] By convention, and in most empirical settings, the threshold k is subject to $T_0(z)$ rather than $T_1(z)$ in a notched schedule.

itive mass in the dominated income region above the threshold (see Saez, 2010; Kleven and Waseem, 2013, for examples). We now turn to a model of frictions that can produce those patterns.

2.2 A sparsity-based model of frictions

We introduce a simple modification to the frictionless model above: rather than selecting from a continuum, taxpayers choose their preferred income from a sparse set of opportunities.

This "sparsity-based" representation of frictions is quite general and can accommodate a diverse set of microfoundations. For example, suppose incomes are produced by performing discrete jobs or gigs that are discovered via a search process. Then the income opportunity set can be interpreted as the incomes available from the set of jobs (or combinations of jobs) that a taxpayer faces after searching with a given intensity. Or a worker in a given position may need to choose between a discrete set of shifts or overtime opportunities, rather than adjusting their labor hours continuously. Alternatively, the model can be viewed as arising from a model of rational inattention, in which a taxpayer can learn the precise income that arises from each potential combination of actions (such as taking a specific series of deductions) only by paying information-gathering costs. If the taxpayer has a limited attention budget, i.e., if the information-gathering costs are substantial, then the income opportunity set can be interpreted as the discrete set of incomes arising from the actions that the taxpayer pays costs to learn about. This 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. More broadly, the information-gathering/observation cost microfoundation has been used extensively 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). In our focal context of firm and household tax optimization, the action space could include both real income determinants—e.g., deciding which income-earning opportunities to pursue—or issues of reporting, e.g., a business owner paying attention to the various tax-deductible expenses they have incurred.[9] This model can also be understood to span imperfect targeting models, in which taxpayers target a desired income, and then realize one or more resulting income opportunities that are offset from their target by an unpredictable error term.

One might expect bunching patterns to be highly sensitive to the details of the income op-

[9]Such frictions might also arise due to "lumpiness" in income reporting opportunities, as when a firm targets their income by advancing some of their (lumpy) payments into the current tax year, or claiming some (lumpy) expenditures as deductible, as discussed in Rees-Jones (2018) and covered extensively in the accounting literature, e.g., Kothari, Leone and Wasley (2005).

portunity process, such as the number of income opportunities or the distribution from which they are drawn. In what follows, we will show that a broad set of opportunity processes give rise to bunching patterns that look quite similar and can be well approximated by a parsimonious limiting case of "uniform sparsity," in which the complexities of the opportunity process are distilled into a single sufficient statistic.

To build intuition, we begin with a simple specification of sparsity-based frictions in which each taxpayer faces an income opportunity set consisting of N draws from a specified distribution centered around their preferred frictionless ("target") income. A convenient feature of this model, which extends to the more generalized models we consider below, is that although individual taxpayers respond to tax reforms discontinuously, the *distribution* of taxpayers responds smoothly, and under a linear income tax, the resulting elasticity of taxable income as in the frictionless model. Intuitively, taxpayers who share a type—i.e., those who target the same level of income—realize incomes that are distributed around that target in a predictable way. Because their target income adjusts smoothly in response to tax reforms, the distribution of realized incomes shifts smoothly as well, with the same change in average incomes as in the model without frictions.

Concretely, we begin with a specification of sparsity-based frictions considered in Saez (1999)—the working paper that preceded Saez (2010)—in which the N opportunities are drawn from a uniform distribution of width W centered on each taxpayer's target income $z^*(n)$. Relative to the frictionless model, this version has two additional parameters: the number of income opportunities N, and the width of the uniform distribution W from which they are drawn. Figure 4 displays a range of simulated income densities that arise from such a model. 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 linear underlying ability density and an elasticity of taxable income of $e = 0.3$; see Section 3 for further simulation details.

Figure 4a displays four simulated income densities in which each taxpayer draws $N = 1, 2, 3,$ or 5 income opportunities from a uniform distribution of width $W =$ 50,000 around their target income. When $N = 1$, the bunching mass takes the shape of a rectangular plateau centered around the kink at the bracket threshold k. This plateau is produced by the mass of bunchers who target the income k but then realize an income that is offset from this target by a uniformly distributed shock. When $N = 2$, 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 4a as $N \to \infty$ is the frictionless model illustrated

in Figure 3b, with no diffusion in the bunching mass. However, there is an alternative notion of a limiting case in which diffusion remains well-defined. We begin with an illustrative example, followed by a formal proposition.

Consider the simulation in Figure 4a with $N = 5$, in which agents draw five opportunities from a window of $W = \$50{,}000$ around their target income. Their income choices are effectively determined by the distribution of just two opportunities, the lowest opportunity above their target and the highest opportunity below their target, which dominate all the other (more distant) opportunities. As a result, this opportunity process with $N = 5$ draws from a window of \$50,000 produces very similar behavior to a process with $N = 10$ draws from a window of $W = \$100{,}000$. Both specifications produce similar distributions over the nearest-to-target opportunities, which are uniformly drawn in the vicinity of the target with the same density $N/W = 5/50{,}000 = 10/100{,}000 = 0.0001$. Motivated by this observation, we consider the behavior of the series of income densities that arise when each agent draws N opportunities from a window of width $N \times 10{,}000$ around their target. Figure 4b plots such a series with the same values of N as in Panel (a). The income density with $N = 1$ exhibits a rectangular plateau centered around the kink, though this time with a width of \$10,000. But as N increases—and the width W increases proportionally—the bunching density appears to converge toward a distinctive "tent shape," with a peak at k and high kurtosis.

Figure 4c 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 4b, the distribution from which 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 in proportion to N.[10] As in Figure 4b, the series appears to converge quickly toward a distinctive tent shape as N increases.

The apparent convergence of the series in Figures 4b and 4c motivates two questions: Do the series in Figures 4b and 4c in fact converge to well-defined limiting densities? And if so, is the limit in the two figures the same?

These questions are answered in the affirmative by the following proposition, which states that as $N \to \infty$, both series converge to the same well-defined limit. Moreover, this limit is not specific to uniform or normal opportunity distributions; in fact it is the limit of such a series for *any* differentiable distribution of opportunities with positive density around the income target. To formalize this, note that the opportunity processes underlying Figures 4b and 4c share a

[10] This figure highlights a conceptual linkage between the two models of frictions simulated in Saez (1999), where the "limited menu of effort" model corresponds to Panel (a) with $N > 1$, and the "uncertainty" model corresponds to Panel (c) with $N = 1$.

common structure in which each agent draws an income opportunity set

$$\{z^*(n)+\varepsilon_1, z^*(n)+\varepsilon_2, \ldots, z^*(n)+\varepsilon_N\},$$

where each ε_i is an iid draw from a distribution $F_\varepsilon(x)$. We construct a series in N by defining $F_\varepsilon^N(x) := F_\varepsilon(x/N)$, so that as N increases, the density of opportunities around the target remains fixed while preserving the position of the target income in the distribution of disturbances. We can then show the following proposition.

Proposition 1. *For any distribution $F_\varepsilon(x)$ with positive continuous density at $x = 0$, the income density arising from a model in which each agent draws N income opportunities offset from their target by iid disturbances $\varepsilon_i \sim F_\varepsilon^N$ converges to the same limit as $N \to \infty$.*

The proof, which relies on notation developed in the next section, is presented in Appendix A.1.

The limiting case referenced in Proposition 1 is one in which income opportunities are a Poisson process. Each taxpayer draws an infinite set of income opportunities spanning the number line, and the probability of drawing an opportunity in any \$1 bin is the same. For this reason, we call this model "uniform sparsity." In the examples in Figures 4b and 4c, this probability (the "arrival rate" parameter of the Poisson process) is 0.0001. Equivalently, the expected difference between adjacent income opportunities is 1/0.0001 = \$10,000; we call this expected distance between adjacent opportunities the "lumpiness parameter."

The bunching pattern in the case of uniform sparsity is also plotted in Figure 4b and 4c, and a striking feature is that the series of densities in these figures converge to the uniform sparsity case very quickly as N increases. The simulation with just two income opportunities looks similar to uniform sparsity, and the simulations with $N = 3$ and $N = 5$ are nearly indiscernible.

Panels (d)–(f) of Figure 4 reproduce the simulations in Panels (a)–(c) in the case of a tax notch, where tax liability increases by \$1000 on incomes above the bracket threshold. In Panel (d)—as in Panel (a)—the case with $N = 1$ 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.[11] As the number of opportunities N increases, the mass develops a distinctive shape with diffuse mass to the left of the

[11] In Panel (d), for $N = 1$, 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 $N = 1$, the observed density at any particular income is simply the average of the density of target incomes—which resembles Figure 4d—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.

threshold and a depression to the right. Panel (e) illustrates how this shape evolves with N when the width W is jointly raised to hold fixed the density of opportunity draws around the target incomes. As in Panel (b), the shape converges to the limiting case of the uniform sparsity model. Panel (f) illustrates this convergence when opportunities are drawn from a normal distribution, as in Panel (c). Although convergence is 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 $N = 5$.

Taken together, Proposition 1 and the simulations in Figure 4 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 and high kurtosis of the Poisson process specification in Figure 4b resemble 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) and the diffuse bunching apparent in our empirical setting, as displayed in Figures 1b and 1c. Similarly, the bunching pattern around a notch in Figure 4c matches key empirical features observed in Kleven (2016) and in our Figure 1a, where taxpayers appear to exhibit "notch-like" behavior. Specifically, the Poisson model produces both diffuse mass to the left of the notch and a positive (but depressed) density in the "dominated region" to the right of the notch.[12] These similarities also provide 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 $N = 1$—those cases are not the ones that appear to have the key features of empirical bunching patterns of interest. This is particularly evident in the case of notches, where specifications with $N = 1$ predict an income density that is *continuous* across the notch threshold at k, in contrast to specifications with $N > 1$, 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) and in Figure 1a clearly exhibit such a discontinuity, suggesting they are better represented by a model with $N > 1$, for which the uniform sparsity model performs well.[13]

A second strength of the uniform sparsity model is its parsimony. It distills the many de-

[12] Kleven and Waseem (2013) propose an alternative model that predicts positive mass in the dominated income range, in which a subset of taxpayers are completely insensitive to the presence of the notch due to adjustment or informational frictions. In such a model, the set of taxpayers who do respond to the notch should bunch precisely at the kink, rather than producing diffuse mass.

[13] 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 just under 4 hours, for example. Notably, the bunching patterns in that paper resemble the shape of the $N = 1$ simulation in Figure 4f, suggesting that marathon times are better modeled by an imperfect targeting model with a single draw, rather than a uniform sparsity model with multiple discrete opportunities.

tails underlying particular sparsity-based models of frictions—such as the 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 benefit 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. 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 issue can be avoided, as the choice of target is irrelevant.[14]

Figure 5 further illustrates the behavior of bunching patterns under the uniform sparsity model. The panels vary the elasticity parameter e and the lumpiness parameter, which we denote μ, in the presence of a kink and a notch. Panel (a) plots densities under the baseline parameter values, as well as with lower and higher values of the elasticity e_0. 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_0, which alters the *spread* of the bunching mass around the kink. Panels (c) and (d) plot such simulations in the presence of a notch.

2.3 Characterizing the bunching mass

In order to estimate the two key parameters of the uniform sparsity model—the elasticity e and the lumpiness parameter μ—it will be useful to have an analytic characterization of the income density produced by this model of frictions. Formally, we assume that within each type n there is a continuum of individuals facing different income opportunity sets, thus producing a continuous type-conditional income density, denoted $g(z|n)$. The observed income density $h(z)$ can then be written as the integral across all types of agents who choose to earn a given income z:

$$h(z) = \int_0^\infty g(z|n) f(n) dn. \tag{5}$$

The uniform sparsity assumption allows us to tractably characterize the type-conditional income density $g(z|n)$. Specifically, the type-conditional density at a given income z' is equal to the probability that an n-type agent draws z', multiplied by the probability that z' is optimal

[14]This feature also allows us to sidestep the question of whether taxpayers *anticipate* frictions when selecting their choice of target. In the case of a notch, for example, sophisticated bunchers should target an income distinctly below the threshold k, in order to reduce the probability of drawing income opportunities on the "wrong side" of the notch, whereas naive bunchers would simply target k, as in the Saez (1999) uncertainty model. Since the Poisson process does not depend on a specified target, this issue becomes irrelevant.

for *n conditional on drawing it*, which we denote $\pi(z|n)$.

Under uniform sparsity, the probability of drawing any particular income, including z', is simply $1/\mu$. This insight reduces the problem of characterizing $g(z|n)$ to the problem of characterizing $\pi(z|n)$, which is facilitated by the uniform sparsity assumption. Under uniform sparsity, the probability of drawing m (a positive integer) income opportunities in an interval between any two incomes z_1 and z_2 is given by the Poisson distribution:

$$\frac{\left(\frac{z_2-z_1}{\mu}\right)^m \exp\left[-\frac{z_2-z_1}{\mu}\right]}{m!}. \tag{6}$$

Of relevance for our application, the probability of drawing *zero* incomes in this interval is therefore $\exp\left[-\frac{z_2-z_1}{\mu}\right]$. We can use this formula to compute $\pi(z|n)$, because the probability that an agent chooses income z' (conditional on drawing it as an opportunity) is simply the probability that they draw zero opportunities in the range of incomes that give them higher utility than z'.

This calculation is illustrated in Figures 6 and 7 for agents of type a, for whom the income tax is locally linear in the vicinity of their preferred incomes. (We will turn to the behavior of bunchers around tax bracket thresholds below.) Figure 6a illustrates the budget constraint for a-type agents' constrained optimization problem. Figure 6b illustrates their indirect utility $v(z;a) = u(z - T(z), z; a)$ as a function of income in the vicinity of their frictionless target income $z^*(a)$. Figure 7 illustrates how this indirect utility function can be used to find the type-conditional density $g(z|a)$ at a particular income z' by computing the dominating income range. We define the functions $\underline{Z}(z'|a)$ and $\overline{Z}(z'|a)$ to return the lower and upper income values that give an a-type agent the same utility as income z'. By construction, if z' lies below a's frictionless target $z^*(a)$, as in the illustration, then $\underline{Z}(z'|a) = z'$. Formally, $\underline{Z}(z;n)$ and $\overline{Z}(z;n)$ are defined as the minimum and the maximum, respectively, of the values of Z that implicitly solve the following equation, for a given z':

$$(1-t_0)z' - \frac{n}{1+1/e}\left(\frac{z'}{n}\right)^{1+1/e} = (1-t_0)Z - \frac{n}{1+1/e}\left(\frac{Z}{n}\right)^{1+1/e}. \tag{7}$$

Putting these calculations together, the type-conditional income density is given by the following equation:

$$g(z|n) = \frac{1}{\mu}\exp\left[-\frac{\overline{Z}(z|n) - \underline{Z}(z|n)}{\mu}\right]. \tag{8}$$

When this density is evaluated at n's target income $z^*(n)$, we get $g(z^*(n)|n) = 1/\mu$, reflecting that if the taxpayer happens to draw $z^*(n)$ as an income opportunity, they will choose it with certainty, and thus the density is simply equal to the probability of drawing $z^*(n)$.

We can combine equations (5) and (8) to characterize the income density $h(z)$ for a given

elasticity e, a lumpiness parameter μ, and a specified ability density $f(n)$:

$$h(z) = \int_0^\infty g(z|n) f(n) dn = \frac{1}{\mu} \int_0^\infty \exp\left[\frac{-(\overline{Z}(z|n) - \underline{Z}(z|n))}{\mu} \right] f(n) dn. \tag{9}$$

We discuss the estimation of this model, including the unobserved type density $f(n)$, in Section 2.6 below. First, however, we describe how this model extends to characterize the income density $h(z)$ around a kink or notch in the tax function.

2.4 Diffuse bunching around tax kinks

The only aspect of the preceding calculation that is affected by the presence of a tax kink is the characterization of the dominating income region $\overline{Z}(z'|n) - \underline{Z}(z'|n)$. This requires characterizing agents' indirect utility functions in the presence of a tax kink. Figure 8 illustrates this construction. Panel (a) plots the budget constraint arising from the kinked tax function as a solid line. The figure also extends each linear segment across the bracket threshold as a dashed line, to illustrate the counterfactual budget constraints that would operate if either $T_0(z)$ or $T_1(z)$ applied across all incomes. The optimal continuous income choice for the marginal non-buncher (type b) is displayed, along with the corresponding indifference curve. Panel (b) displays the indirect utility function for the marginal non-buncher, which can be found by retaining the relevant segments of the indirect utility functions that would arise under each of the linear budget constraints:

$$v(z;b) = \begin{cases} v_0(z,b) & \text{if } z \leq k \\ v_1(z,b) & \text{if } z > k \end{cases} \tag{10}$$

Panels (c) and (d) likewise display the budget constraint and the composite indirect utility function for the marginal buncher, type c. In both cases, the kink in the budget constraint produces a corresponding kink in agents' indirect utility functions.

Figure 9 illustrates how these indirect utility functions are translated into type-conditional income densities for the marginal non-buncher (Panel (a)) and the marginal buncher (Panel (b)). Both panels illustrate the calculation of the type-conditional income density at an income z'—which differs across the panels—by identifying the dominating income range for z' as perceived by each type. We then proceed as in the case of a linear tax, already described, by computing the type-conditional income density using equation (8).

The result of this type-conditional density calculation is plotted in the lower portion of each panel in Figure 9. In the case of the marginal non-buncher (Panel a), the effect is to raise the type-conditional density at z', relative to the density $g_0(z|b)$ that would obtain under the linear tax $T_0(z)$, which is plotted for comparison. The reason for this change can be understood from

the size of the dominating income range above: some incomes above the threshold k that would be preferred to z' under the linear tax $T_0(z)$ are no longer preferred in the presence of a kink (i.e., $\overline{Z}(z';b)$ is lower than in the absence of the kink). As a result, it is less likely that the taxpayer would draw another income that dominates z', and thus the probability that one chooses z' (conditional on drawing z') has increased. Applying this logic to other income choices, we can trace out the shape of the type-conditional density $g(z|b)$ across all incomes. The result is that the kink-induced type-conditional density is left skewed relative to the counterfactual $g_0(z|b)$. By the same logic, the kink-induced type-conditional density for type c has right skew.

The type-conditional densities illustrated in Figure 9 cannot be observed in the data because types are unobservable. Figure 10 illustrates the implications for the observable density of *incomes*. The top portion of the figure shows the optimal continuous choice for agents of types a, b, c, and d in the presence of a kink. The lower portion illustrates the overlapping type-conditional income densities of each type. (Although we have plotted these densities for only four types, the underlying type distribution is continuous across this range by assumption, so that there is a continuum of types between these four.) The observable income density, labeled $h(z)$, results from "adding up" the type-conditional densities in Panel (a), and the other intermediate unplotted types. The clustering of the type-dependent densities of b and c around the kink produces diffuse excess mass in the income density $h(z)$ around the kink at k. If the underlying type density is uniform, as in this example, the excess mass is approximately symmetric because the asymmetry in the type-conditional income densities of b and c balances out, as illustrated by Figure 9. A sloped type density will tend to produce excess mass that is not perfectly symmetric, although it is still of a form that can be fully computed for a given type density $f(n)$ and parameters e and μ.

2.5 Asymmetric bunching around tax notches

The model above also extends to the case of notches, with one additional nuance: notches may produce non-monotonicities in the indirect utility function. Figure 11 illustrates the construction of the composite indirect utility functions for the marginal non-buncher and buncher. In each case, the discontinuity in the budget constraints in Panels (a) and (c) produce corresponding discontinuities in the indirect utility functions in Panels (b) and (d). As shown in Panel (d), the marginal buncher has two local maxima in indirect utility: k, and the maximal point of $v_1(z;c)$.

To construct the type-conditional income densities, we proceed as before by computing the range of dominating incomes at any given z'; this procedure is illustrated in Figure 12. In the case of the marginal non-buncher (type b), this calculation is straightforward; we need only

modify the calculation of $\overline{Z}(z';n)$ to reflect the possibility that the dominating region may be bounded above by k, as illustrated in Panel (a). In the case of the marginal buncher (type c), the calculation is complicated by existence of two different income ranges that dominate z', below and above the threshold k. To handle such cases, we define the functions $\underline{Z}_1(z;n)$ and $\overline{Z}_1(z;n)$ to return the lower and upper incomes that produce utility equal to $v(z;n)$ under the linear income tax $T_1(z)$, and $\underline{Z}_0(z;n)$ to return the lower income value that produces utility equal to $v(z;n)$ under the linear tax $T_0(z)$. Thus in Panel (b), the left dominating income range is the interval $\left[\underline{Z}_0(z';n),k\right] = \left[z',k\right]$, and the right dominating income range is the interval $\left[\underline{Z}_1(z';n),\overline{Z}_1(z';n)\right]$. A virtue of the uniform sparsity model is that the probability of drawing zero opportunities in a dominating range does not depend on the position of the range with respect to the frictionless target income, nor on its contiguity. The probability depends only on the size of the dominating income range(s) in total. Therefore, once these dominating income ranges are determined, they can simply be summed, and the result inserted into the numerator of the bracketed term in equation (8) to compute the type-conditional density.

Figure 13 illustrates the aggregation of the type-conditional income densities to construct the observed income density. As in Figure 10, the top panes of each panel display the income choices of the four types a, b, c, and d. Summing these type-conditional densities (as well as those of all the unplotted intervening types) produces the observed density $h(z)$, plotted in the lower portion of the figure. This density exhibits the key features observed in empirical settings with tax notches: diffusion in the excess mass left of the threshold, and positive mass in the dominated income region to the right.

2.6 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 in the rest of the bunching literature, 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, smooth in the vicinity of the bracket threshold k. Intuitively, this amounts to assuming that the location of the bracket threshold is not selected to occur at an income that happens to coincide with a distortion in the underlying ability distribution.[15]

[15] Blomquist et al. (2021) explore this identification strategy and its limitations at length. See Moore (2022) for a

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

$$f(n;\theta) = \theta_0 + \theta_1 n + \theta_2 n^2 + \dots \tag{11}$$

$$= \sum_{q=0}^{Q} \theta_q n^q \tag{12}$$

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

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

$$L(e, \mu, dT, \theta) = \Pi_i h(X_i = z; e, \mu, dT, \theta). \tag{13}$$

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

$$\max_{e,\mu,\theta,dT} \sum_i \log h(X_i = z; e, \mu, dT, \theta) \quad \text{subject to} \quad \int_{z_{min}}^{z_{max}} h(z; e, \mu, dT, \theta)\, dz = 1. \tag{14}$$

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 (14) 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 (14) 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 (14) presents

discussion of what can be identified by bunching estimators without estimating the elasticity directly.

a challenge. The likelihood function is

$$h(z;e,\mu,dT,\theta) = \int_{-\infty}^{\infty} g(z|n;e,\mu,dT) f(n;\theta) dn. \tag{15}$$

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

$$\max_{e,\mu,dT} \sum_i H_i \log h(Z_i = z;e,\mu,\theta(e,\mu,dT)), \tag{16}$$

with the integration constraint in (14) 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 (14), 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:

$$\theta(e,\mu,dT) = \min_{\theta} \sum_i \left(\frac{H_i}{\sum_j H_j} - h(Z_i;e,\mu,dT) \right)^2. \tag{17}$$

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 $\sum_i \varepsilon_i^2$:

$$\begin{aligned}
\frac{H_i}{\sum_j H_j} &= h(Z_i;e,\mu,dT) + \varepsilon_i \\
&= \int_{-\infty}^{\infty} g(Z_i|n;e,\mu,dT) f(n;\theta) dn + \varepsilon_i \\
&= \int_{-\infty}^{\infty} g(Z_i|n;e,\mu,dT) \left(\theta_0 + \theta_1 n + \theta_2 n^2 + \theta_3 n^3\right) dn + \varepsilon_i \\
&= \left[\int_{-\infty}^{\infty} g(Z_i|n;e,\mu,dT) dn\right] \theta_0 + \left[\int_{-\infty}^{\infty} g(Z_i|n;e,\mu,dT) n dn\right] \theta_1 \\
&\quad + \left[\int_{-\infty}^{\infty} g(Z_i|n;e,\mu,dT) n^2 dn\right] \theta_2 + \left[\int_{-\infty}^{\infty} g(Z_i|n;e,\mu,dT) n^3 dn\right] \theta_3 + \varepsilon_i.
\end{aligned} \tag{18}$$

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 (16), searching over only the three parameters e, μ, and dT. The integration constraint in equation (14) can be enforced by using a two-step procedure, in which after selecting a provisional θ vector to solve equation (17), we adjust the intercept θ_0 so that the constraint holds exactly.

In spirit, this method resembles the approach—often employed in the conventional bunching literature—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 a window around the threshold to exclude when computing the best-fit values of θ. Instead, even data near the bracket threshold helps identify θ. This reasoning suggests that this estimation method may be more robust to choices about the polynomial degree Q than is the conventional bunching estimator, where additional flexibility may attempt to fit excess mass that spills outside the excluded bunching window. We confirm this reasoning in the 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 3, 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 A.5.

We specify a simulated tax kink using the same parameters as in Figures 4 and 5: 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 impose a 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 1. (The simulations in Figures 4 and 5 used a much higher population size of 2 million to illustrate the shape of the bunching mass with less sampling noise.)

We construct these simulated 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.[16] 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.[17]

3.1 Performance of our estimator and the standard 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 obtain joint estimates of $\hat{e}$ and $\hat{\mu}$. We are interested in whether the distribution of these estimates is centered around the true parameter values 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 14a. 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

[16] 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 $z^*(\overline{n}) = \overline{z} + 100{,}000$.

[17] 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 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.

visual inspection, we employ the algorithmic method preferred by Bosch, Dekker and Strohmaier (2020). The details of this implementation are described in Appendix A.2. Figure 14b presents the results. The bunching window is bounded by dashed lines, and the orange line displays the fitted counterfactual density outside that window.[18] 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 15a 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 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 15b. To construct this figure, we reproduce distributions like those in Figure 14a using several different values of the lumpiness parameter μ_0. Figure 15b 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

[18] 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.

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 15, we note that this source of bias should become more severe when the polynomial fit is allowed to be more flexible. In Appendix Figure A1, we reproduce the estimates in Figure 14 with different polynomial degrees of 1 (linear), 3, 5, and 10. Consistent with this story, Figure A1b 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 A1a 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 14a 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.[19]

[19]Describing the rationale for imposing the integration constraint, Chetty et al. (2011) remarks that an unad-

However, as illustrated by the hollow blue points in Figure 3a, 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.[20]

To explore the role of the integration constraint, Appendix A.4 reproduces the results in Figure A1b 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 A.2 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 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,

justed 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."

[20] Indeed, Figure 3b, 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.

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 A.5, 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.[21]

4 Empirical application

We apply our estimation method using administrative data on the income distribution of firms around the three prominent tax kinks in the Small Business Corporation tax schedule in South Africa. The bunching patterns at each kink are displayed in Figure 1, and the underlying schedule of marginal tax rates is described below.

4.1 Data and background on 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, this tax base accounted for 16.2 percent of total tax revenue in South Africa, considerably higher than the OECD average of 9.3 percent and in line with the average share for Africa (18.6 percent) and Latin America (15.3 percent).[22] The South

[21] 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 5.)

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

African corporate income tax consists of a tiered system, with a progressive, kinked tax schedule applying to "Small Business Corporations" (SBCs), and a flat 28 percent tax applying to other resident companies.[23] Corporate taxable income consists of gross revenues less non-capital expenses and less any incurred losses from previous tax years which can be carried forward.[24] There are no local corporate income taxes in South Africa; businesses pay income tax only at the national level.[25]

We focus on SBCs because their kinked tax schedule is a natural setting for the bunching estimation approach. 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).[26] We describe the full set of eligibility requirements, and other details of SBCs, in Appendix A.6. 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 tax revenue.

Most relevant for our application, SBCs face a piecewise-linear progressive kinked tax schedule. The lowest threshold, at which the marginal tax rate rises from 0 to 7 percent, is at R75,750, or about $5,260 US. (In 2018, South African GDP per capita was about $7,000 US.) Below this threshold, firms face no tax liability, although they are still legally required to file a tax return. This threshold moves over time with inflation. 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 so that firms with incomes above this threshold face the same marginal tax rate as non-SBC firms. Appendix Figure A5 plots the schedule for 2018, 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 A5 has been in place.

4.2 Results

We now estimate the uniform sparsity model of frictions around each of the three tax kinks in the South African SBC schedule. The results are plotted in Figure 16, Panels (a)–(c). The empirical histogram of firms around each kink are displayed in green, and the model-predicted densities produced by our maximum likelihood estimates are shown in orange. Parameter es-

[23] There are also alternative tax schedules for gold mining companies and micro businesses, neither of which are the focus of this paper.

[24] Corporate dividends are taxed at the shareholder level, at a 15 percent rate.

[25] See Pieterse, Gavin and Kreuser (2018) for more on the South African corporate income tax data. We plan to analyze personal tax returns in future research.

[26] 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.

timates for the income elasticity (e), the lumpiness parameter (μ), and the "as-if" notch value (dT), together with their 95 percent confidence intervals, are reported in the upper right corner of each plot. Although these tax thresholds correspond to statutory kinks, we allow the model to flexibly estimate the notch value for reasons discussed below.

We find income elasticity estimates of 0.27 and 0.23 at the middle and upper kinks, respectively. The estimates are not statistically distinguishable ($p > 0.05$). The elasticity estimated at the lowest kink is substantially higher, in excess of one. This is perhaps not surprising, as the base level of income is much lower at this kink. (In the extreme, as incomes approach zero, any measurable behavioral response would correspond to a high elasticity.)

Panels (a)–(c) of Figure 16 also report the estimated lumpiness parameters at each kink, which range from R5,900 to R11,300 ($410 to $785 US). These estimates provide insight into the extent of income frictions faced by firms at each tax kink. Between the lower and middle kinks, μ appears to increase with income, as would be the case if the distances between lumpy income opportunities increase with one's total income. However, the estimates are non-monotonic, declining from the middle to the upper kink. This pattern may be explained by heterogeneity in tax practitioner usage, discussed below.

We compare our elasticity estimates to those produced by the conventional kink-based bunching estimator, computed using the method described in Appendix A.2. These estimates are reported in Panels (d)–(f) of Figure 16. Each panel plots the estimated counterfactual density in orange.[27] These elasticity estimates are substantially lower, falling below the maximum-likelihood-based elasticity estimates by between 30 and 50 percent. 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 imprecision of the conventional bunching estimator may be economically important in practice.

In addition to providing estimates of the elasticity and lumpiness parameters, our estimation method uncovers additional insights about the economic behavior of businesses in this setting. As noted in the introduction, although the bunching patterns around the middle and upper thresholds in Figure 16 have the visual features typically associated with a tax kink, the lower kink exhibits strikingly asymmetric bunching, with a discontinuity at the bracket threshold and missing mass to the right, which is suggestive of the bunching patterns around a notch. This raises the question of whether businesses are behaving "as if" there is a notch at the lowest threshold. Our estimation method provides a framework for answering this question formally. In Figure 16a, the estimated notch value is R340, or about $24 US, and is highly statistically significant, suggesting that the model strongly rejects pure kink behavior. Interestingly, the es-

[27] 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.

timated notch value at the middle kink is similar in magnitude and also statistically significantly different from zero. This result highlights that conditional on a given notch value, the degree of visual asymmetry is heavily mediated by the income elasticity—an insight consistent with the simulations displayed in Figure 5c. 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 rather be a more general phenomenon. As such, it suggests that the source of this "as-if" notch value is unlikely to be driven solely by a behavioral aversion to paying a positive tax liability. Although identifying the source of "as-if" notch behavior is beyond the scope of this paper, 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. 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 additionally explore heterogeneity in bunching behavior across firm attributes. Specifically, we compare businesses that use a registered tax practitioner to those that do not. Figure 17 displays plots like those in Figure 16, partitioning businesses on tax practitioner usage. The raw histograms exhibit notably more pronounced bunching among firms with tax practitioners. Table 1a reports these results, together with the estimates for the aggregate population in Figure 16. 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. However, a clear and consistent difference can be discerned in the lumpiness parameters within each group. At every kink, income frictions appear to be smaller among firms who use tax practitioners. This would be consistent with such firms fine-tuning their incomes more precisely in response to tax incentives, or paying closer attention to the various actions—whether related to real economic activity or reporting behaviors—which can be used to target their incomes to a desired level.[28]

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 other firms. If "as-if" notch behavior is driven by average-versus-marginal tax rate confusion of the nature described above, this result would be consistent with tax-practitioner-using firms exhibiting less such confusion. This result may help explain the small size of the estimated notch value at the upper kink in the aggregate sam-

[28]These estimates may help explain the non-monotonicity in μ across incomes in Figure 16. After conditioning 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 may not be economically meaningful.

ple (Figure 16c), which appears to be driven primarily by businesses with paid tax practitioners.

This heterogeneity in behavior across tax practitioner usage, although visually apparent in Figure 17, is undetected by the conventional bunching estimator approach. Table 1b reports the income elasticities estimated by the conventional approach at each of the three kinks. Naturally, the conventional bunching estimator does not provide insights into differences in the degree of income frictions or the "as-if" notch value, since these parameters are not estimated by that model. However, it also fails to detect differences in the parameter that it does measure—the income elasticity—which are evident under the maximum likelihood approach. The behavior around the middle kink illustrates this phenomenon. Comparing Panels (b) and (e) in Figure 17, firms with tax preparers exhibit a substantially—and statistically significantly—lower income elasticity than those without tax preparers (0.26 vs. 0.50). But as reported in Table 1b, the conventional bunching estimator returns statistically indistinguishable elasticity estimates (0.13 vs. 0.11). In words, income frictions may lead the conventional approach to misinterpret heterogeneity in lumpiness or "as-if" notch behavior in a way that masks real differences in the elasticity parameter of interest.

5 Conclusion

This paper extends the theory underlying bunching-based elasticity estimators to incorporate a positive model of frictions, and provides new estimation methods to recover these elasticities. We consider a general class of sparsity-based frictions in which taxpayers select their preferred option from a sparse set of opportunities. This setup can be microfounded using a range of models of frictions, 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.

We conduct simulations generated assuming sparsity-based frictions, which suggest that when applying the standard bunching estimator, sparsity-based frictions confound the estimation of the counterfactual income density. This can lead to underestimating both the bunching mass and the elasticity, while producing overly precise confidence intervals. We propose an alternative estimation method that recovers unbiased elasticity estimates in simulated data. This approach can be used to draw novel economic insights by quantifying the extent of frictions and estimating the "as-if" notch value when taxpayers treat a kink like a notch. We apply this method to administrative tax data on small firms around three prominent tax kinks in South

Africa, where we estimate income elasticities between 0.2 and 0.3 at the middle and upper tax kinks and high elasticities in excess of one at the lowest kink. We find evidence that taxpayers 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.

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, forthcoming* .

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.

Dekker, Vincent and Karsten Schweikert. 2021. "A Comparison of Different Data-driven Procedures to Determine the Bunching Window." *Public Finance Review* 49(2):262–293.

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* .

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.

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.

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.

Mavrokonstantis, Panos and Arthur Seibold. 2022. "Bunching and Adjustment Costs: Evidence from Cypriot Tax Reforms." *Working Paper no. 9773, CESifo* .

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.

Pollinger, Stefan. 2021. "Kinks Know More: Policy Evaluation Beyond Bunching with an Application to Solar Subsidies." *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.

Saez, Emmanuel, Joel Slemrod and Seth H Giertz. 2012. "The elasticity of taxable income with respect to marginal tax rates: A critical review." *Journal of Economic Literature* 50(1):3–50.

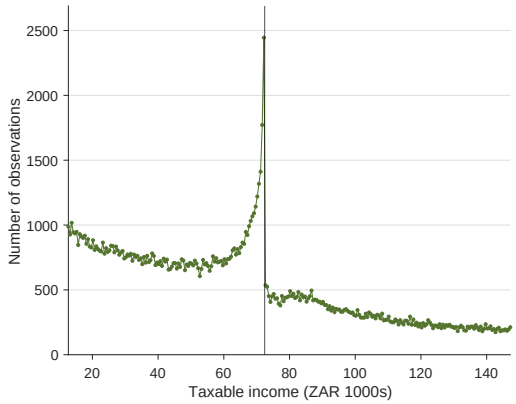
Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50(3):665–690.

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

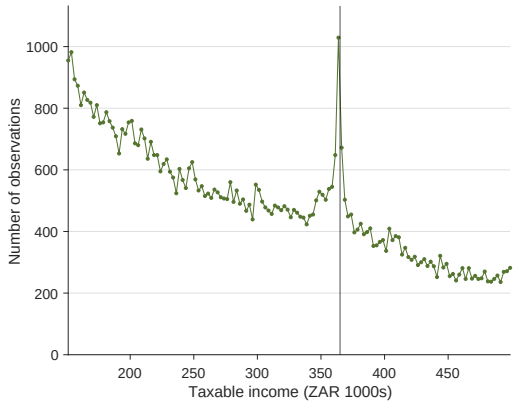
Velayudhan, Tejaswi. 2018. Misallocation or misreporting? evidence from a value added tax notch in india. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*. Vol. 111 JSTOR pp. 1–46.

Figure 1: Bunching in the income distribution of South African Small Business Corporations

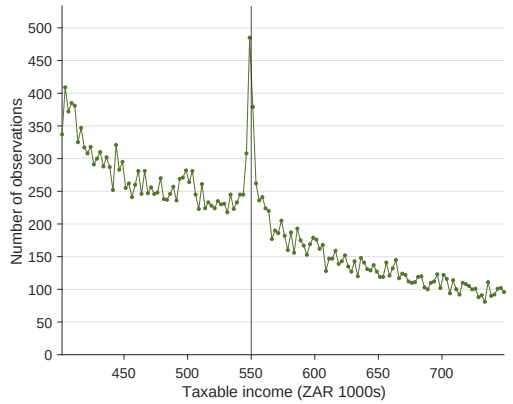
(a) Lower kink



(b) Middle kink

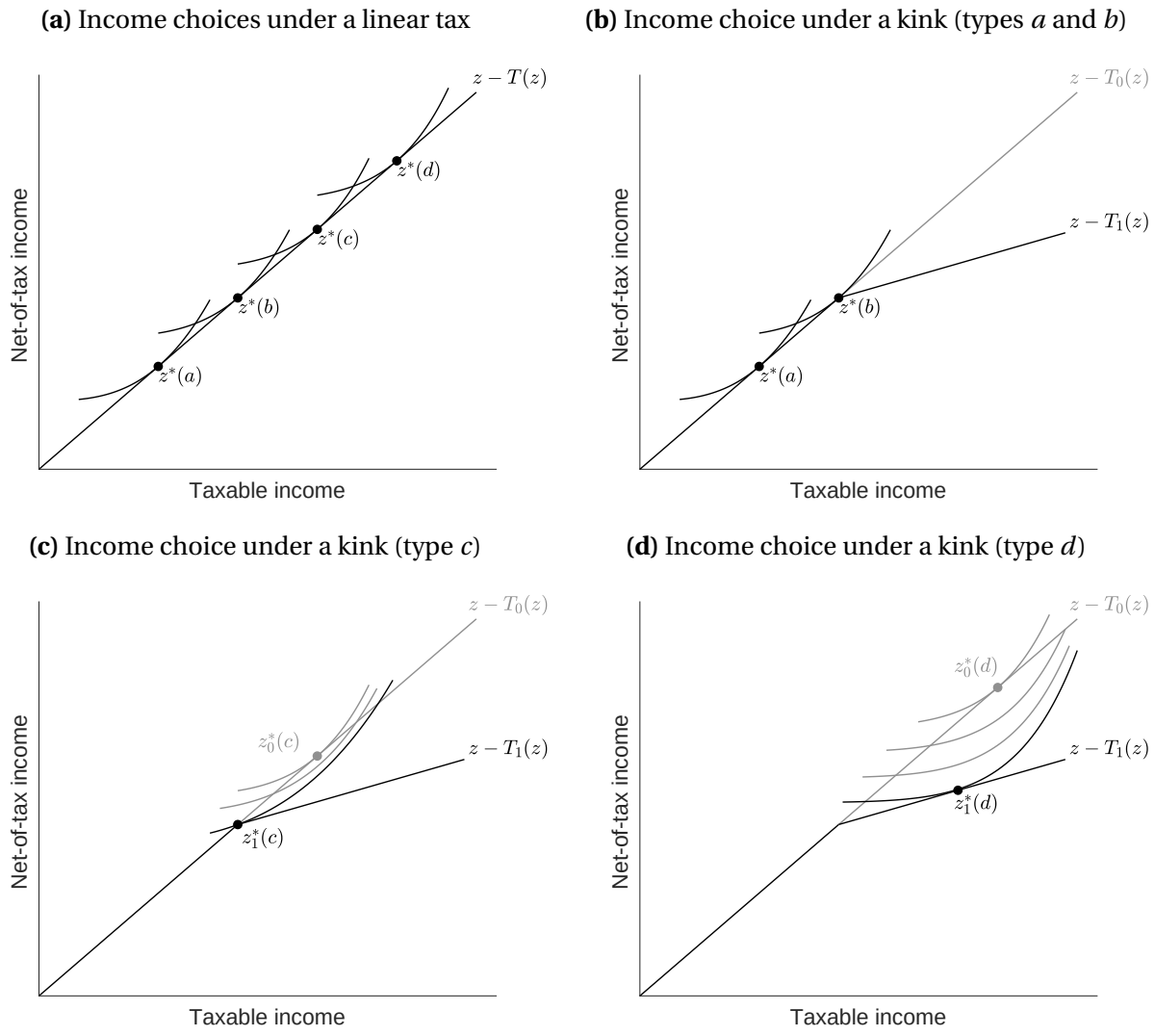


(c) Upper kink



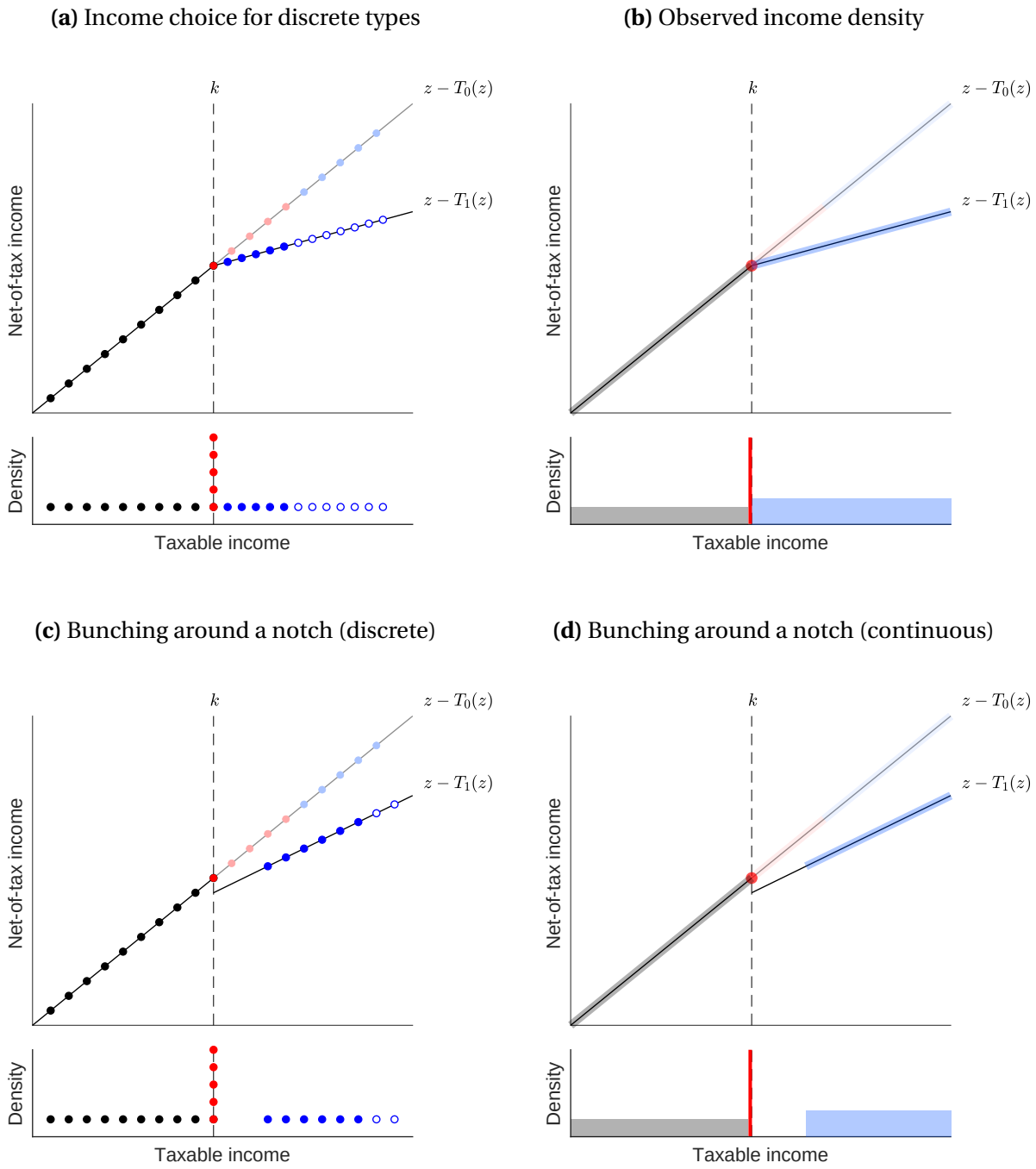
The green points plot the empirical histogram of firms with different earnings in the data. The sample consists of all South African small business corporations that filed tax returns in 2014–2018. The lowest bracket threshold adjusts over time for inflation. In this and all similar figures, we shift the histogram along the lower threshold within each year so that each year's threshold aligns at the same point; the horizontal axis reflects the lower bracket threshold in 2018.

Figure 2: Conventional (continuous-choice) 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.

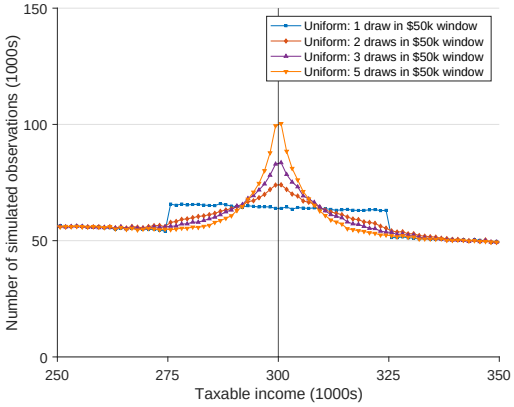
Figure 3: Bunching patterns under a model without income frictions



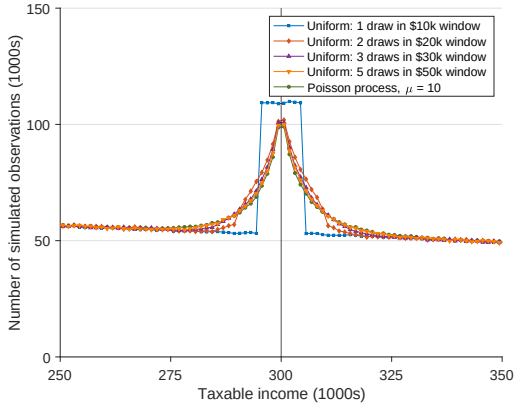
Panel (a) plots income choices for discrete types in the presence of a kink, for a uniform type distribution. Black points denote types who choose the same income under the linear tax $T_0(z)$ and the kinked tax schedule. Red points denote types who bunch at the bracket threshold k under the kinked tax 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 4: Simulated bunching patterns with sparsity-based frictions

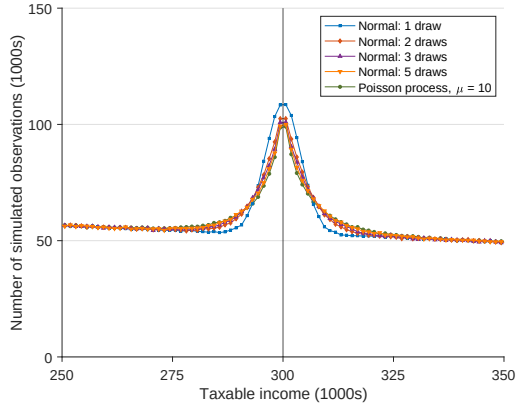
(a) Uniform draws, fixed support (kink)



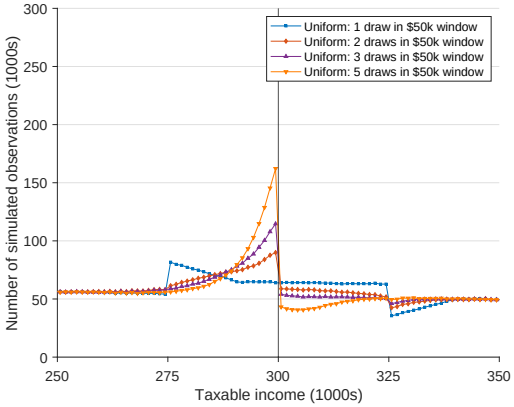
(b) Uniform draws, fixed density (kink)



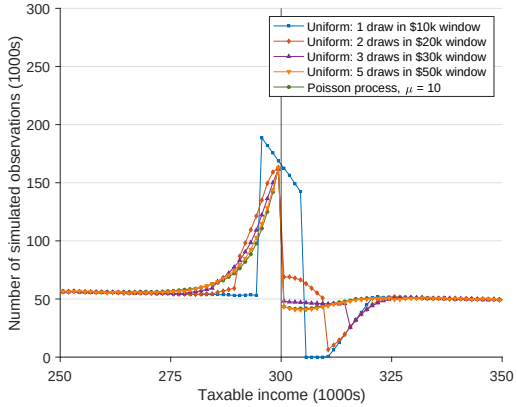
(c) Normal draws, fixed density (kink)



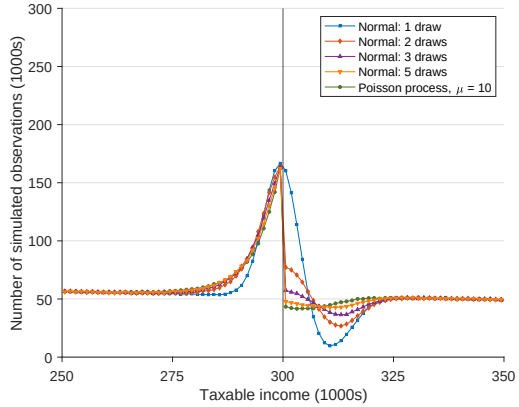
(d) Uniform draws, fixed support (notch)



(e) Uniform draws, fixed density (notch)



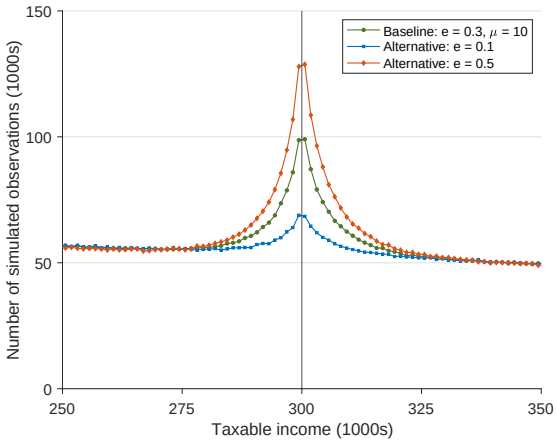
(f) Normal draws, fixed density (notch)



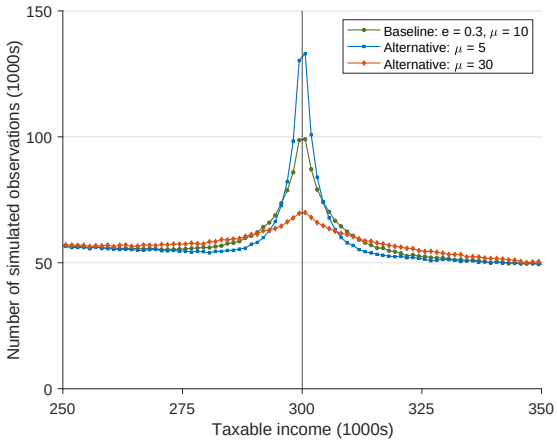
This figure plots simulated income densities around a bracket threshold under a model of frictions in which each taxpayer faces a sparse set of income opportunities drawn from around their preferred frictionless ("target") income. In all simulations, the marginal tax rate rises from $t_0 = 0.1$ to $t_1 = 0.2$ 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 N income opportunities drawn from a uniform distribution of width $50,000 around their target income, for $N = 1, 2, 3$, and 5. In Panels (b) and (e), each taxpayer chooses from N 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 $N \to \infty$—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 5: Simulated income densities under the uniform sparsity model

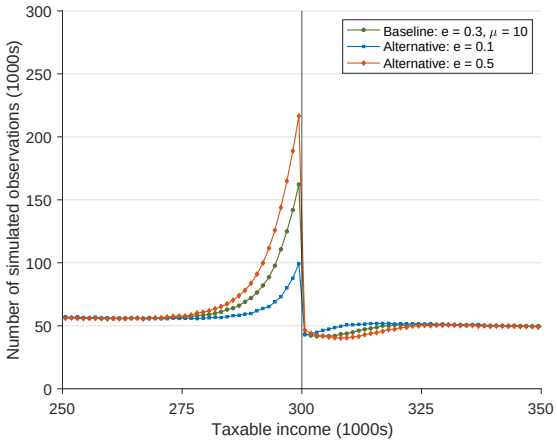
(a) Different elasticities (kink)



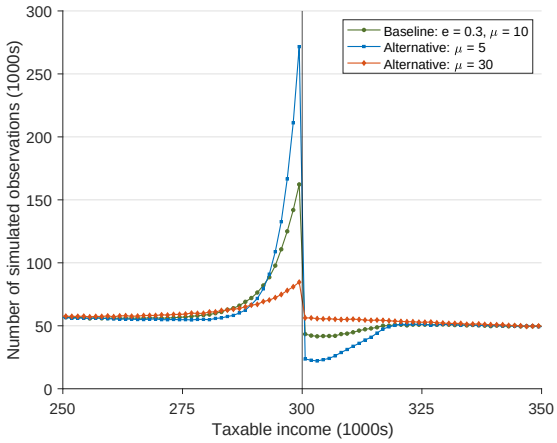
(b) Different lumpiness parameters (kink)



(c) Different elasticities (notch)

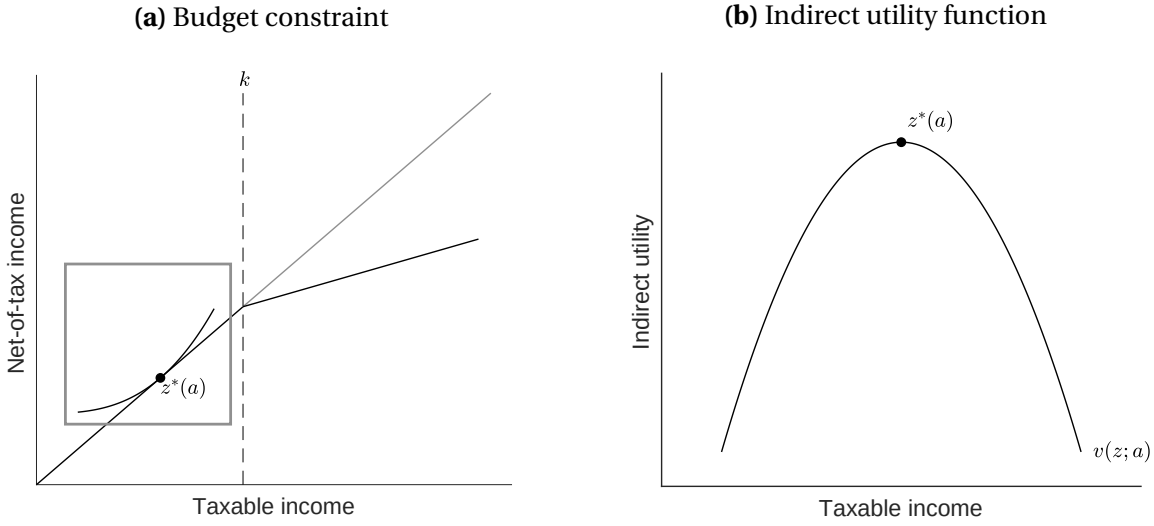


(d) Different lumpiness parameters (notch)



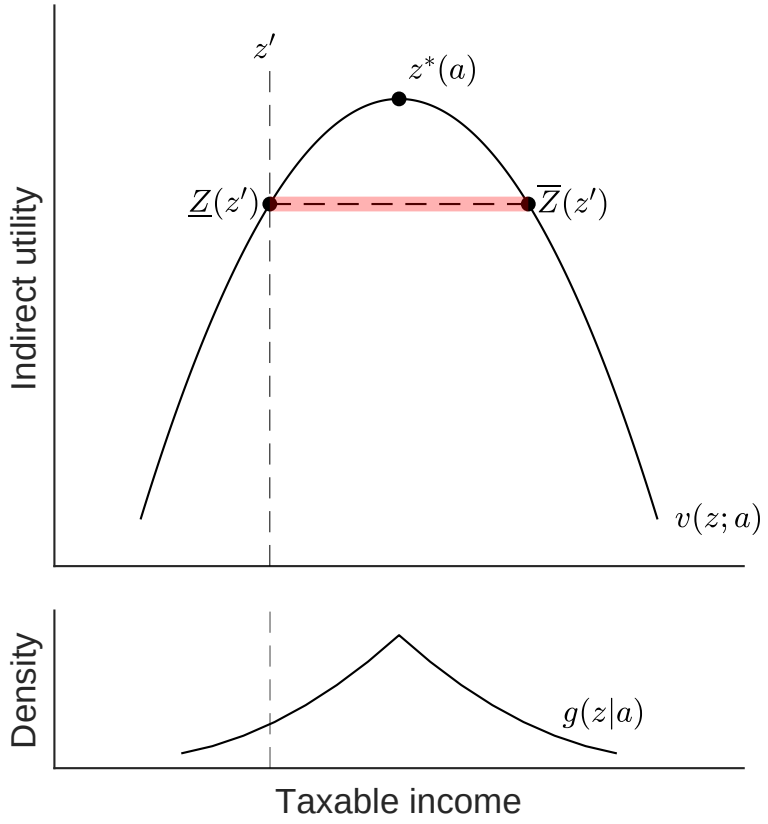
This figure plots income histograms from simulated data sets under the uniform sparsity model of frictions. For each simulation, we draw agents from an ability distribution with a linear density. We assume agents have a homogeneous income elasticity, e_0, and for each agent we then draw a sparse set 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 6: Utility from income choices under a linear tax (type a)



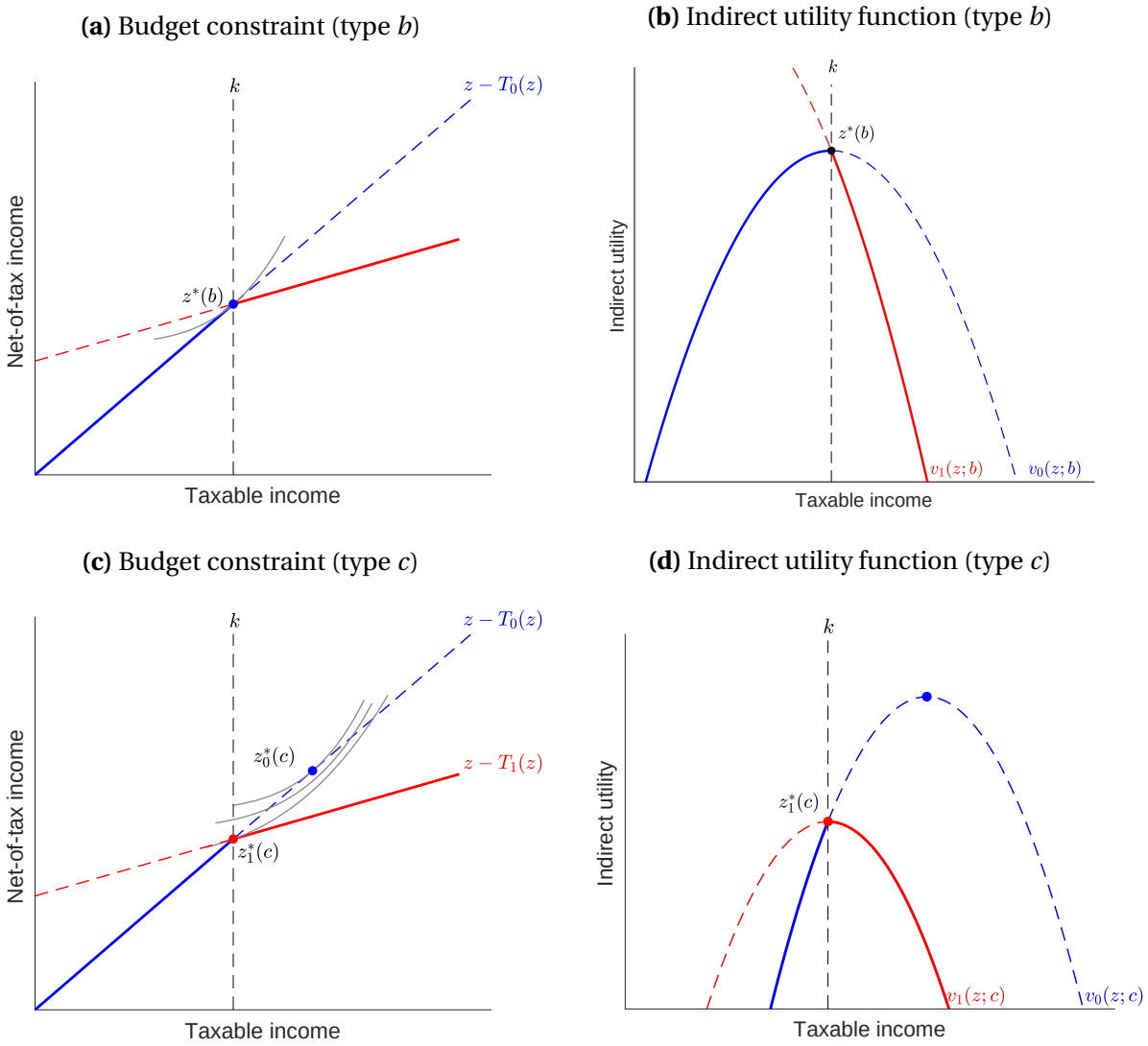
Panel (a) shows the optimal choice of continuous income for an agent type a. Panel (b) plots this type's indirect utility function $v(z;a) \equiv u(z - T(z), z; a)$ in the neighborhood of this optimal choice.

Figure 7: Type-conditional income density under a linear tax (type a)



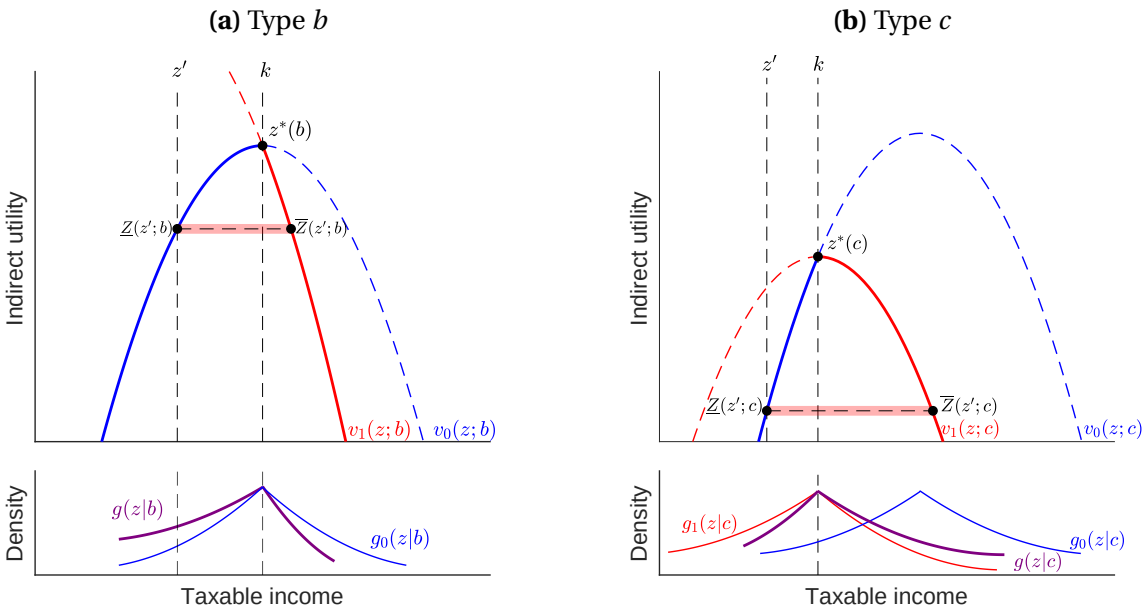
This figure illustrates the calculation of the type-conditional income density in the uniform sparsity model among a-type agents at a particular income level z', under a locally linear income tax. The top portion reproduces the indirect utility function from Figure 6b. There is a continuum of a-type agents facing this indirect utility, each of whom draws a sparse set of income opportunities from a Poisson process with lumpiness parameter μ. An agent who has z' in their income opportunity set will select this income iff they do not have some other income opportunity that yields higher utility, i.e., iff they do not have an income opportunity in the "dominating income range" between $\underline{Z}(z')$ and $\overline{Z}(z')$ in the figure above. The probability of having zero income choices in this interval, given by the Poisson distribution, is $\pi(z'|a) = \exp\left[\frac{-(\overline{Z}(z'|a)-\underline{Z}(z'|a))}{\mu}\right]$. The type-conditional density $g(z'|a)$ is equal to this conditional probability multiplied by the probability of drawing z', which is $1/\mu$.

Figure 8: Utility from income choices around a tax kink



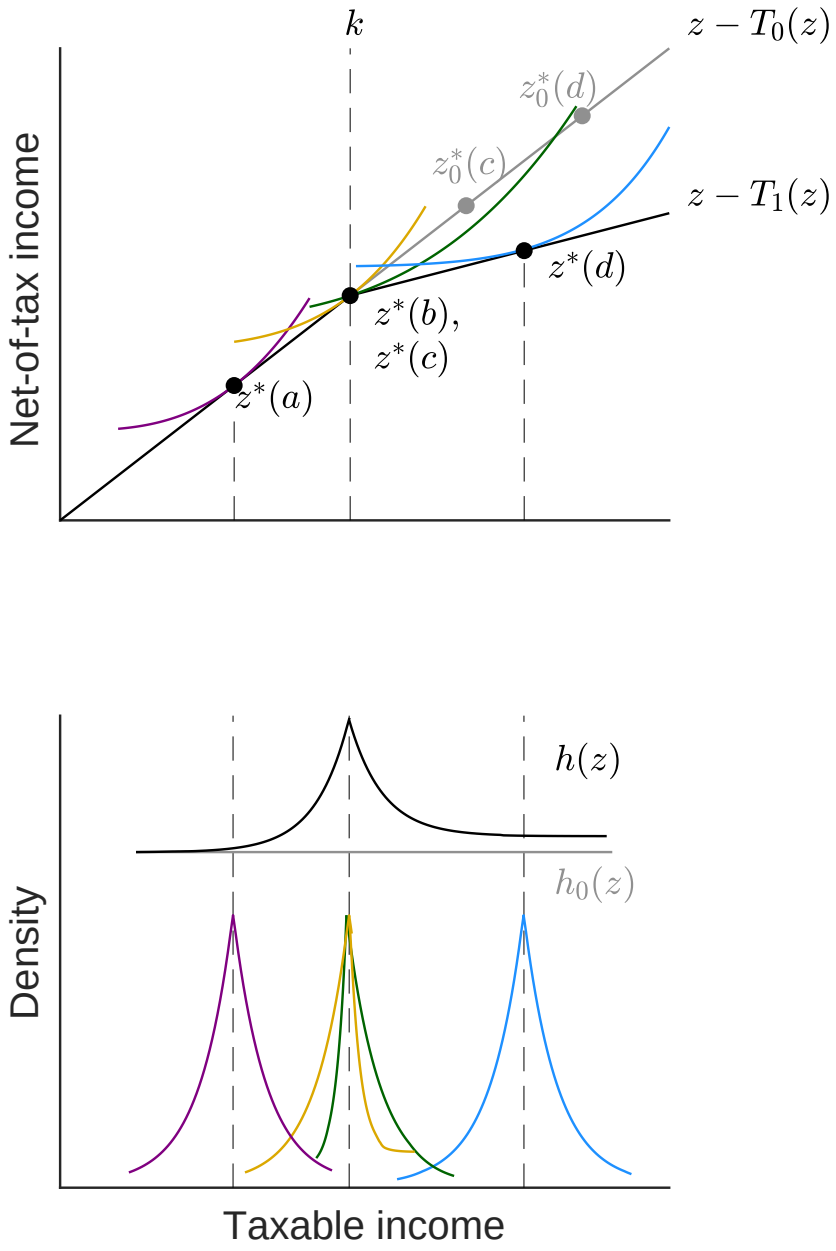
Panels (a) and (b) illustrate the construction of the indirect utility function around a progressive tax kink for the marginal non-buncher (type b). Panel (a) shows the taxpayer's budget constraint, plotted as a solid line, where $T_0(z)$ and $T_1(z)$ are the linear income taxes below and above the bracket threshold k, respectively. Panel (b) 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 (c) 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 9: Type-conditional income density around a kink



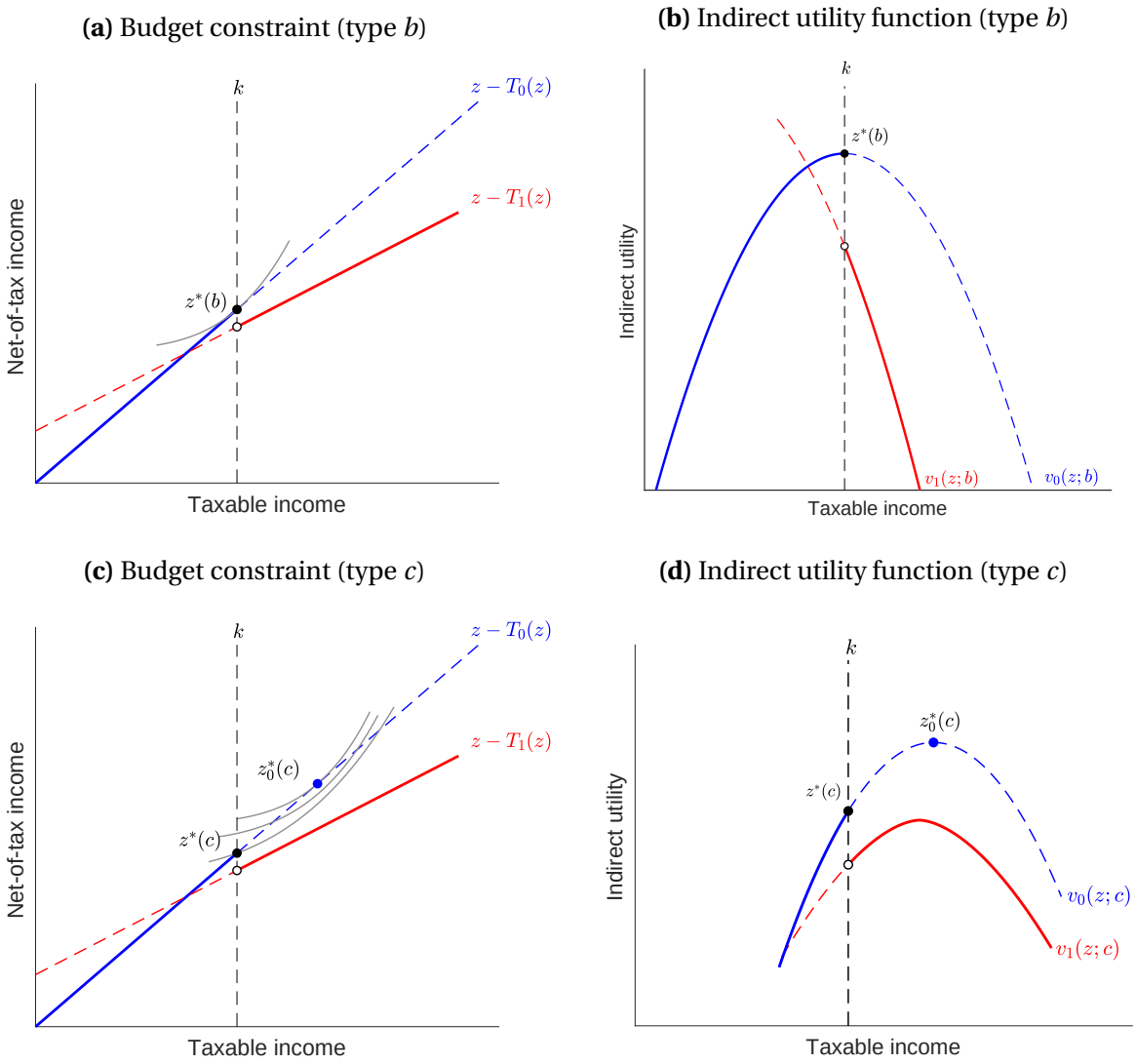
This figure illustrates how the indirect utility functions from Figure 8 are used to compute the type-conditional income densities. The panels show the calculation for the marginal non-buncher (Panel (a)) and the marginal buncher (Panel (b)). Each panel illustrates the calculation of the type-conditional income density $g(z|n)$ at a (different) income z'. We first identify the range of incomes that dominate z' for each taxpayer, corresponding to the horizontal dashed line, and we proceed as in Figure 7. 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 10: Aggregating type-conditional densities into observable income density (kink)



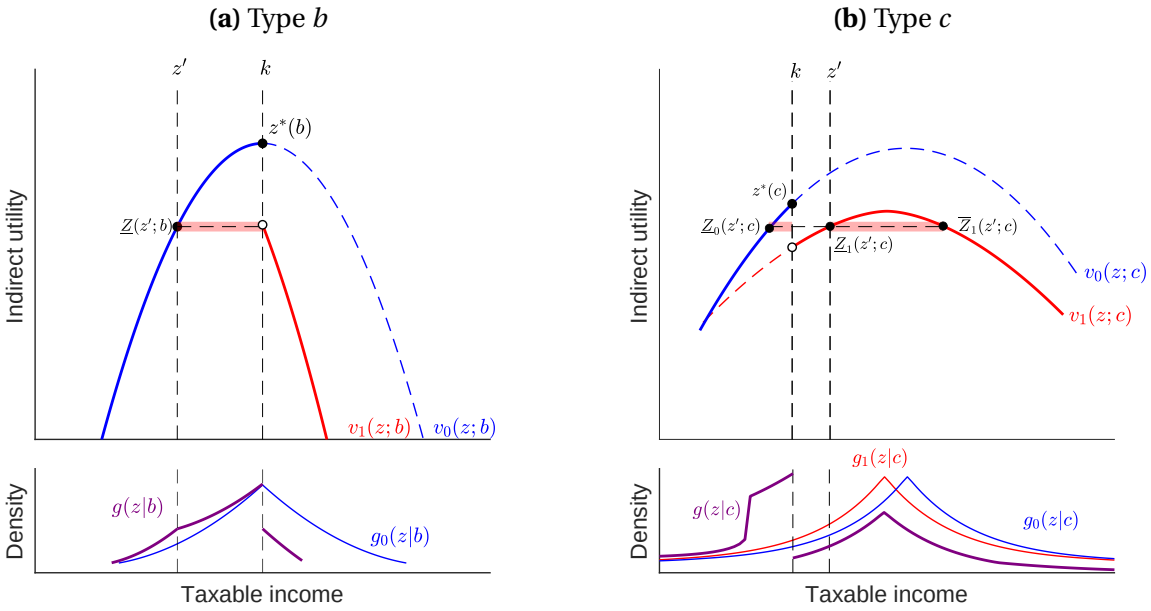
The top portion of this figure shows the optimal frictionless income choice for agents of types a, b, c, and d in the presence of a kink at k, with each type's maximal indifference curve plotted in a different color. The lower portion plots the type-conditional income densities in corresponding colors. Summing across the type-conditional densities of these and 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 apply under the linear tax function $T_0(z)$, is plotted in gray for reference.

Figure 11: Utility from income choices around a tax notch



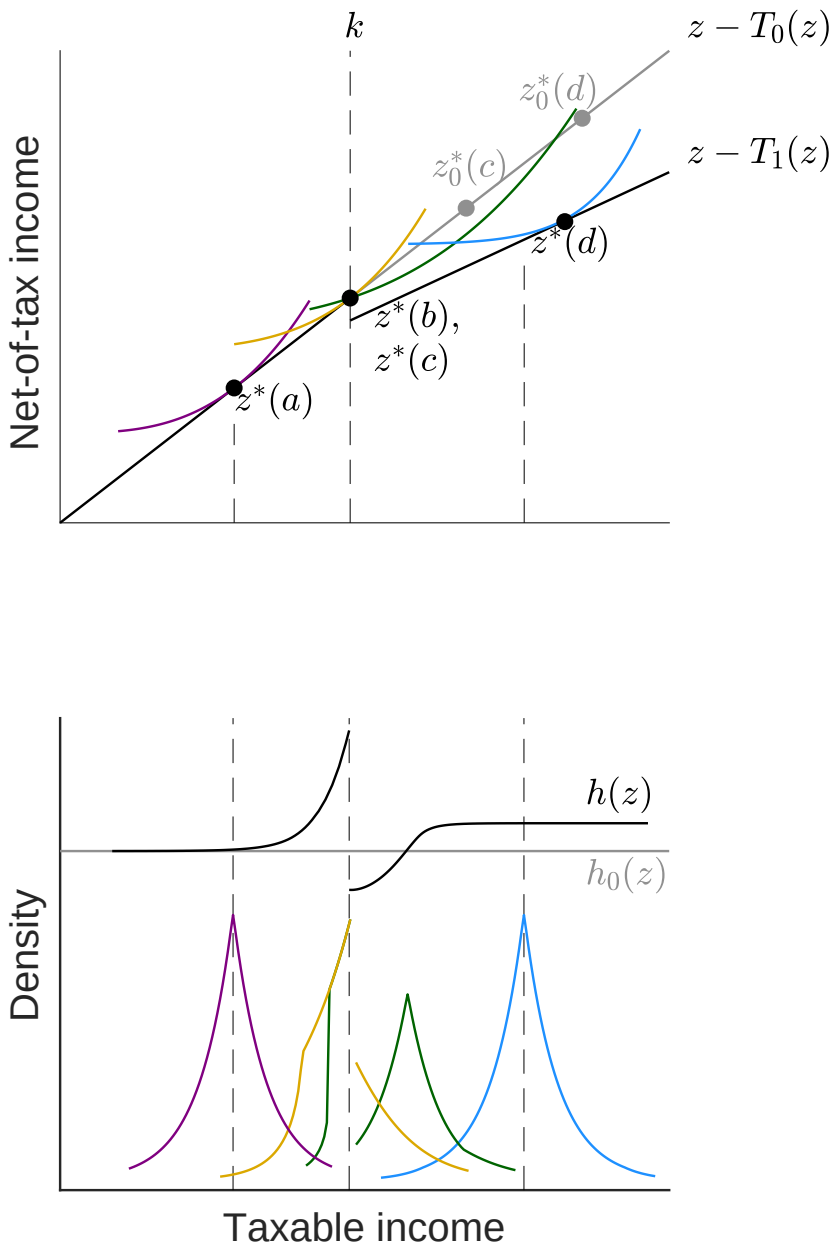
This figure is analogous to Figure 8, but in the presence of a notch, which produces a discontinuity in the indirect utility function (Panels (b) and (d)). In the case of type c, the notch produces a non-monotonic indirect utility function with two local maxima (Panel (d)).

Figure 12: Type-conditional income density around a notch



This figure is analogous to Figure 9, 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.

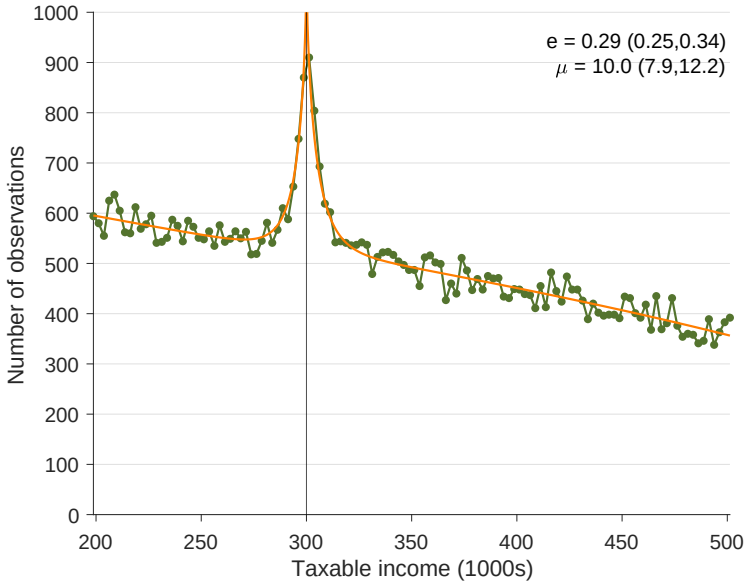
Figure 13: Aggregating type-conditional densities into observable income density (notch)



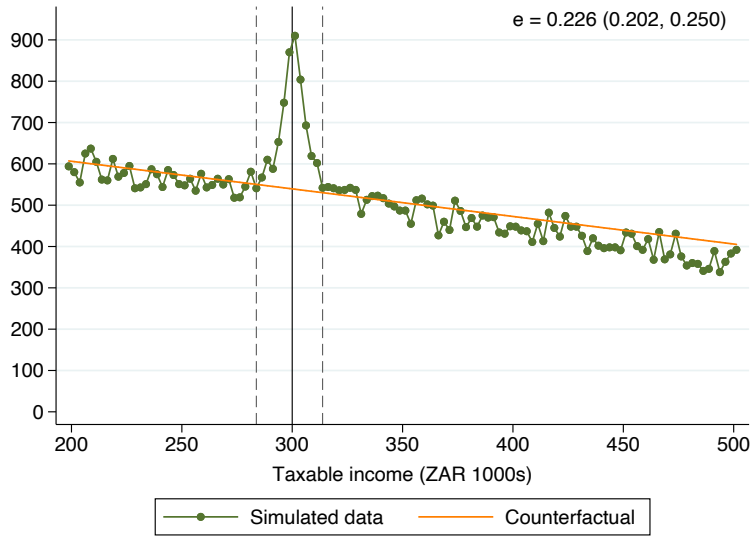
This figure is analogous to Figure 10, but in the presence of a notch. The asymmetric excess mass to the left of the threshold k accumulates across types, producing asymmetry in the observed income density $h(z)$.

Figure 14: Parameter estimates from simulated data

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



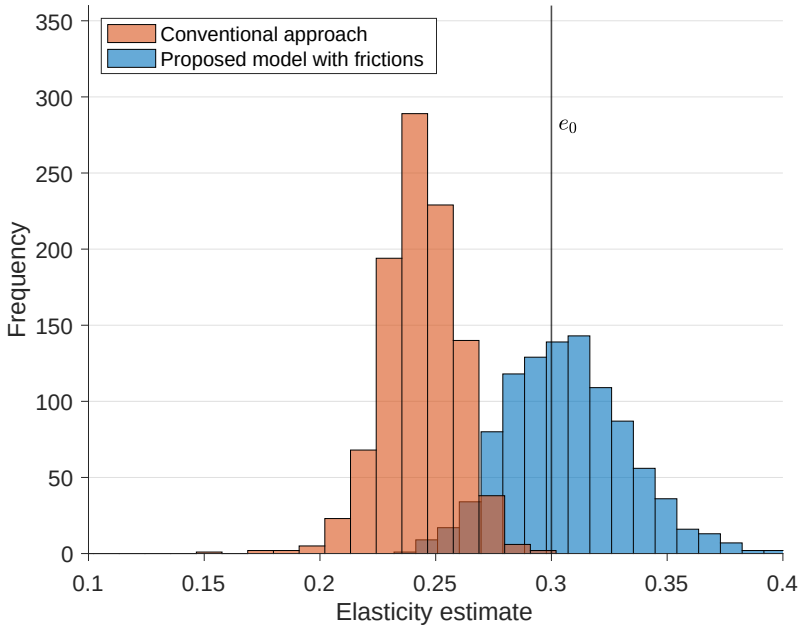
(b) Conventional bunching estimator for a single simulation round



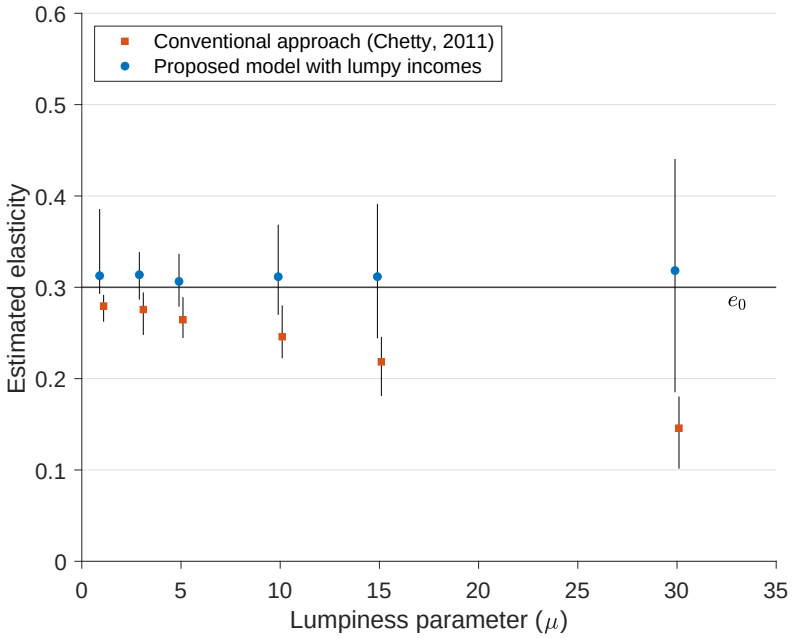
This figure displays the estimation of the maximum likelihood model and the conventional bunching estimator for one round of simulated data. The simulation is constructed as in Figure 5, with a true elasticity of $e_0 = 0.3$ and a lumpiness parameter of $\mu_0 = 10$, but using a smaller number of drawn observations ($N = 100{,}000$) 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 2.6. 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 15: Elasticity estimates using the conventional approach

(a) Distribution of elasticity estimates under each approach

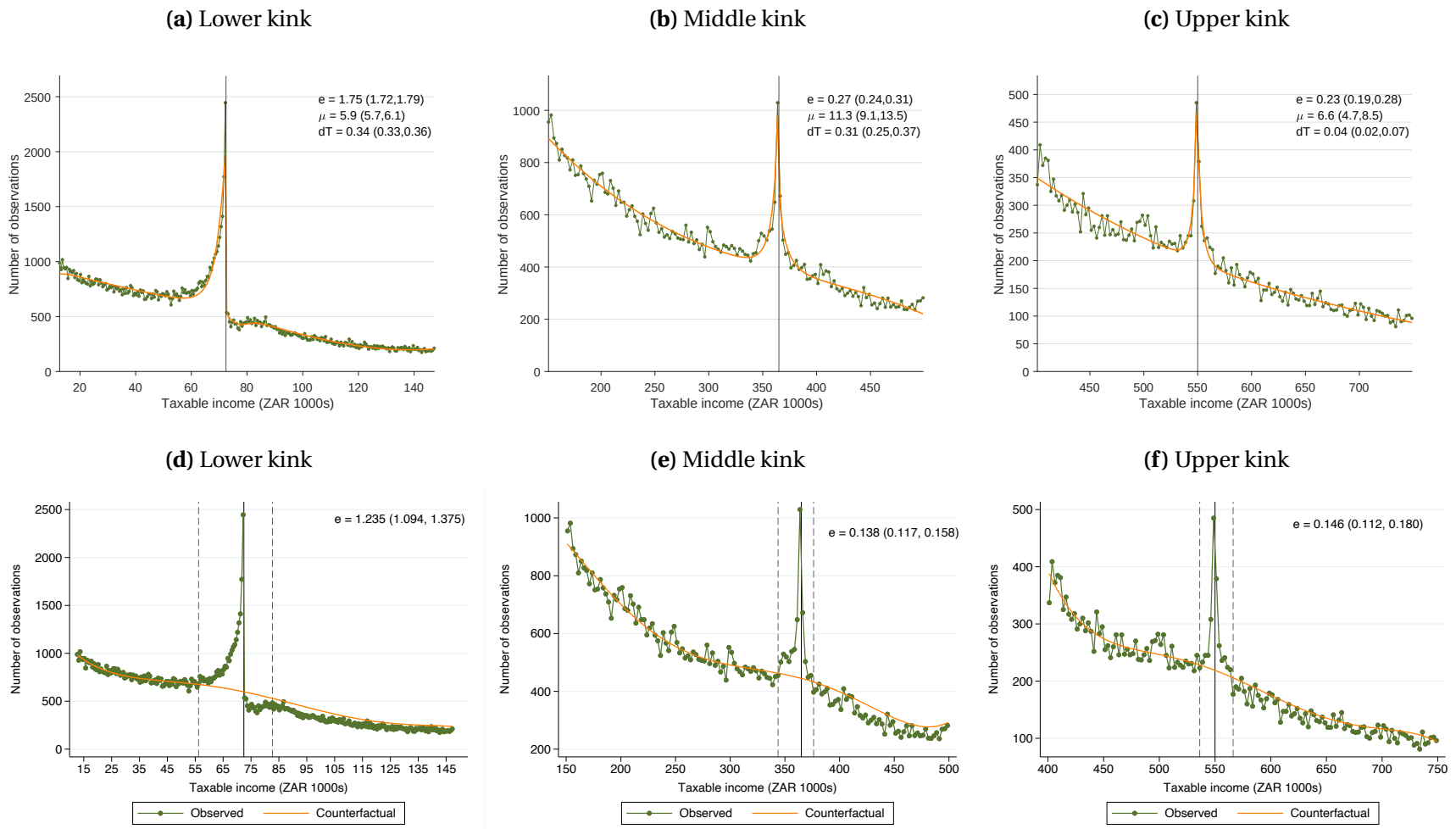


(b) Elasticity estimates under each approach for different lumpiness parameters



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.

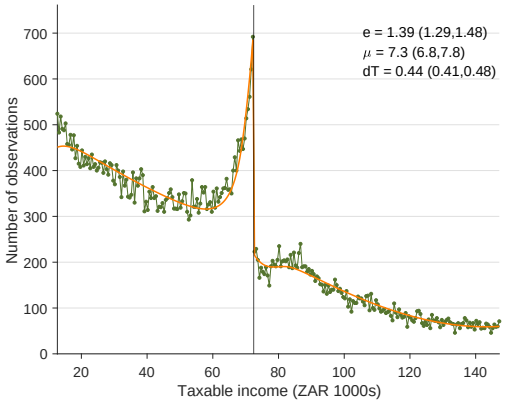
Figure 16: Empirical application model estimates



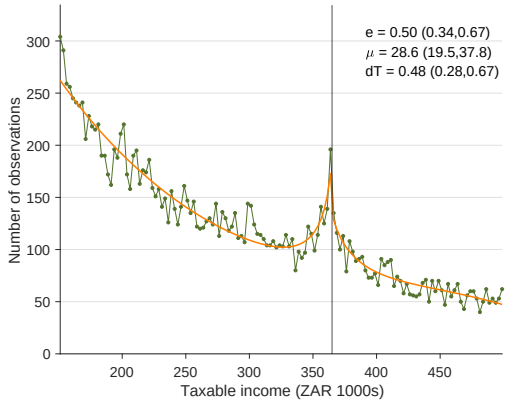
Green points plot the empirical histogram of firms with different earnings in the data. In Panels (a)–(c), orange lines plots the predicted density generated by the maximum likelihood estimation of the uniform sparsity model parameters e (elasticity of taxable income), μ (average distance between income opportunities in ZAR 1000s), and dT (the estimated "as-if" discrete change in tax liability at the bracket threshold, in ZAR 1000s). In Panels (d)–(f), the orange line plots the predicted density generated by fitting a polynomial counterfactual to the histogram 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 A.2. 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 17: Heterogeneity by tax practitioner usage

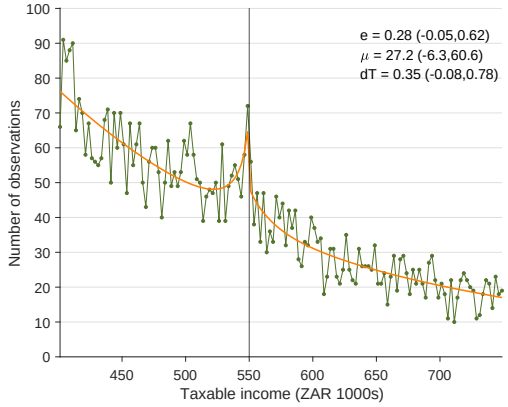
(a) No tax practitioner, lower kink



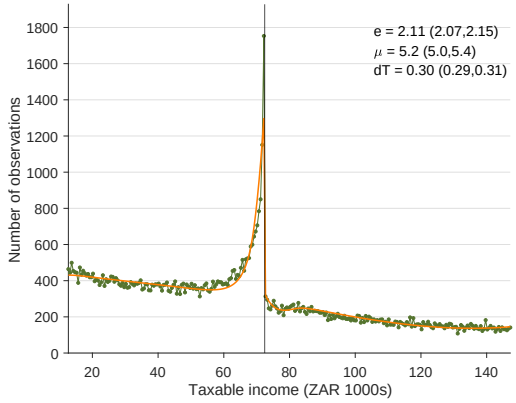
(b) No tax practitioner, middle kink



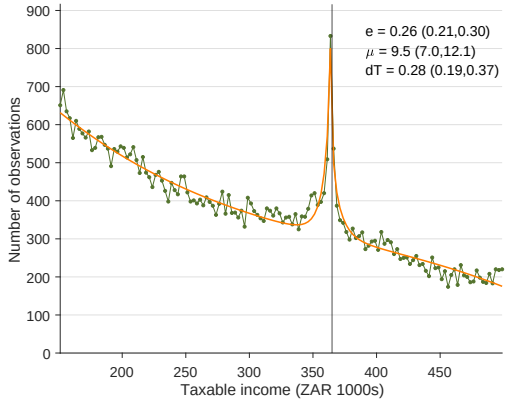
(c) No tax practitioner, upper kink



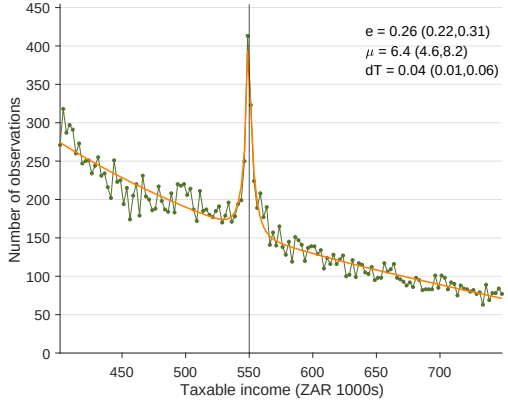
(d) Uses tax practitioner, lower kink



(e) Uses tax practitioner, middle kink



(f) Uses tax practitioner, upper kink



The format of these plots is the same as in Panels (a)–(c) of Figure 16, but the sample is split into firms that do and do not use professional tax practitioners to prepare their tax returns.

Table 1: Estimated bunching parameters

(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 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 A.2.

A Online Appendix

A.1 Proof of Proposition 1

The simulations in Figures 4b and 4c share a common structure. In both cases, agents draw an opportunity set consisting of N opportunities

$$\{z_1, z_2, \ldots, z_N\} = \left\{z^* + \varepsilon_1, z^* + \varepsilon_2, \ldots, z^* + \varepsilon_N\right\}, \tag{19}$$

where the ε are iid random draws from a particular distribution—uniform in 4b and normal in 4c—which we can more generally denote F_ε, with density f_ε. 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 (since all higher income opportunities are dominated by that draw) and the highest negative ε (since all lower income opportunities are dominated by that draw). We can view these two draws as a pair of order statistics, of a sort. Importantly, the ultimate income density will depend solely on the distribution of these order statistics. The number of draws N, 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 $\varepsilon_j > 0$, the set of incomes that are less desirable are all those $z^* + \varepsilon_k$ with 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, for simplicity let $\underline{\phi}(\varepsilon)$ and $\bar{\phi}(\varepsilon)$ denote the functions that compute the utility-equivalent disturbances from z^* that deliver the same utility as a given disturbance ε_j, so that $\underline{Z}(z^* + \varepsilon) - z^* = \underline{\phi}(\varepsilon)$ and $\bar{Z}(z^* + \varepsilon) - z^* = \bar{\phi}(\varepsilon)$ for all ε.

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

$$N f_\varepsilon(\varepsilon_j) \left[1 - (F_\varepsilon(\varepsilon_j) - F_\varepsilon(\underline{\phi}(\varepsilon_j))\right]^{N-1} \tag{20}$$

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

$$N f_\varepsilon(\varepsilon_j) \left[1 - (F_\varepsilon(\bar{\phi}(\varepsilon_j)) - F_\varepsilon(\varepsilon_j)\right]^{N-1} \tag{21}$$

in the case of $\varepsilon_j < 0$. Together, these equations characterize the *type-conditional income density*

at a given income, which we write as

$$g(z^*(n)+x|n) := Nf_\varepsilon(x)\left[1-(F_\varepsilon(\bar{\phi}(x))-F_\varepsilon(\underline{\phi}(x))\right]^{N-1}. \tag{22}$$

This characterizes the probability that agent's 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 $\bar{\phi}(\varepsilon_j)=\varepsilon_j$ when $\varepsilon_j>0$ by construction to combine the preceding equations.

Evaluating equation (22) at $x=0$ gives the local density of opportunity draws around the agent's target income z^*, which is $Nf_\varepsilon(0)$. Figures 4b and 4c and the informal accompanying discussion suggest that this density is a key determinant of the patterns of bunching around tax kinks. Those figures illustrate this phenomenon by selecting a parametric form for the distribution F_ε (either uniform or normal) and then adjusting that distribution to hold fixed the opportunity density $Nf_\varepsilon(0)$ while raising N, from which we observe apparent convergence as $N\to\infty$.

Here, we formalize and generalize that argument. We allow F_ε to be any distribution with a continuously differentiable density and with $F_\varepsilon(0)>0$, so that the density of income opportunities around the agent's income target is positive. As in the heuristic argument beside figures 4b and 4c, we wish to adjust the number of opportunities N while jointly adjusting the distribution F_ε to hold fixed the local density of opportunities around the target income, $Nf_\varepsilon(0)$, as N increases. In figures 4b and 4c, this transformation entailed widening the support of the uniform distribution (4b) or increasing the variance of the normal distribution (4c) so as to hold fixed the density of income opportunities around z^*. Generally, such transformations, indexed by N, can be defined for an arbitrary distribution F_ε as follows:

$$F_\varepsilon^N(x) := F_\varepsilon\left(x/N\right), \tag{23}$$

which implies

$$f_\varepsilon^N(x) = \frac{1}{N} f_\varepsilon(x/N). \tag{24}$$

We call $F_\varepsilon^1(x)=F_\varepsilon(x)$ the *unitary distribution.* As N increases, this transformation holds fixed $Nf_\varepsilon^N(0)$, the density of opportunity draws around the target income. In keeping with the lumpiness parameter we define in the text, we will call this constant value $1/\mu$. The transformation also preserves $F_\varepsilon^N(0)$, meaning that the target income remains at the same quantile in the distribution from which opportunities are drawn. Substituting F_ε^N and f_ε^N into equation (22), we can compute the type-conditional income density as a function of N:

Regarding F_ε^N and f_ε^N as any particular distributions from which N opportunities are drawn,

equation (22) provides the type-conditional income density, which we now index by N:

$$g^N(z^*(n)+x|n) := Nf_\varepsilon^N(x)\left[1-(F_\varepsilon^N(\bar{\phi}(x))-F_\varepsilon^N(\underline{\phi}(x))\right]^{N-1}. \tag{25}$$

Using equations (23) and (24), this can be rewritten in terms of the unitary distribution:

$$g^N(z^*(n)+x|n) = f_\varepsilon(x/N)\left[1-\left(F_\varepsilon\left(\frac{\bar{\phi}(x)}{N}\right)-F_\varepsilon\left(\frac{\underline{\phi}(x)}{N}\right)\right)\right]^{N-1}. \tag{26}$$

We are interested in the limit of this function $g^N(z|n)$ as $N\to\infty$. It is:

$$g^\infty(z^*(n)+x|n) = f_\varepsilon(0)\left[1-\left(F_\varepsilon\left(\frac{\bar{\phi}(x)}{N}\right)-F_\varepsilon\left(\frac{\underline{\phi}(x)}{N}\right)\right)\right]^{N-1} \tag{27}$$

$$= \frac{1}{\mu}\left[1-\frac{\bar{\phi}(x)-\underline{\phi}(x)}{\mu}\cdot\frac{1}{N}\right]^{N-1} \tag{28}$$

$$= \frac{1}{\mu}\exp\left[-\frac{\bar{\phi}(x)-\underline{\phi}(x)}{\mu}\right] \tag{29}$$

The final line follows from $\lim_{N\to\infty}(1-x/N)^N = e^{-x}$, the definition of continuous exponential decay. Using the definitions of $\underline{\phi}$ and $\bar{\phi}$, we can express this the type-conditional density among types n at any income z:

$$g^\infty(z|n) = \frac{1}{\mu}\exp\left[-\frac{\bar{Z}(z)-\underline{Z}(z)}{\mu}\right].$$

Going forward, we simplify notation by writing this limiting case simply as $g(z|n)$.

A.2 Details of the conventional kink-based bunching estimator

We apply the conventional bunching estimator based on Saez (2010) to estimate the income elasticity the simulated data sets underlying Figure 14. 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 A.4 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 specifi-

cation:

$$C_j = \sum_{i=0}^{q} \beta_i^0 \cdot (Z_j)^i + \sum_{i=R_l}^{R_u} \gamma_i^0 \cdot \mathbf{1}[Z_j = i] + \epsilon_j^0. \tag{30}$$

Here, q denotes the order of the polynomial, and R_l and R_u denote the lower and upper bounds of "bunching window" near the kink, which is excluded from the polynomial estimation.[29] 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.[30]

The second step is to compute the excess mass of incomes around the kink relative to this counterfactual density. Using equation (30), 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:

$$\hat{B} = \sum_{i=R_l}^{R_u} C_j - \hat{C_j^0} = \sum_{i=R_l}^{R_u} \hat{\gamma}^0. \tag{31}$$

We then map this excess mass estimate to an estimated elasticity using the approximation from Chetty et al. (2011):

$$\hat{e} \approx \frac{\hat{B}}{z^* \cdot h_0(z^*) \cdot \log(\frac{1-t_0}{1-t_1})} \tag{32}$$

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.[31]

[29] 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.

[30] 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 A.4, our results confirm that the integration constraint introduces bias, and that the introduced bias may be substantial.

[31] This approach proceeds in five steps: (1) Estimate equation (30) with no bunching window—so that the poly-

A.3 Additional figures from simulation estimates

Figure A1 reproduces the estimations in Figure 14 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.

Figure A2a plots the joint distribution of estimates $(\hat{e}, \hat{\mu})$ for the 1000 simulation rounds underlying Figure 14a. 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 $\mu_0 = 10$.

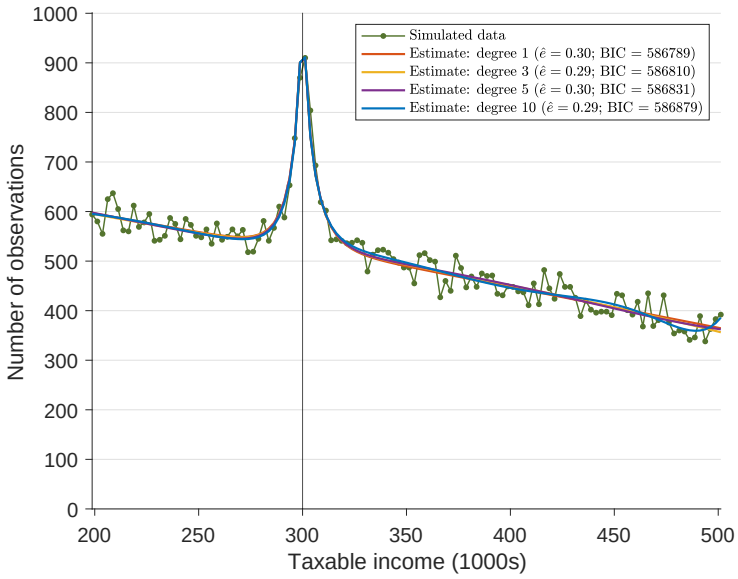
Second, the spread of both distributions provides an indication of sampling error. In each round of simulated data, the maximum likelihood estimation procedure also provides a standard error estimate, and so a key question is whether this estimate gives an accurate picture of the degree of precision 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 14a 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 14b 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 14b.

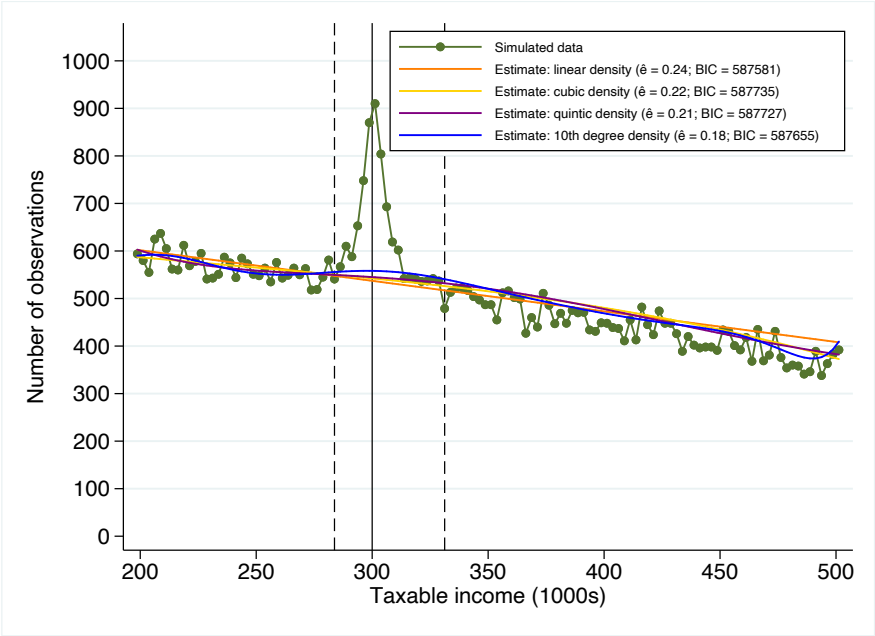
nomial estimation excludes only the bins adjacent to the kink—for a range of polynomial orders, retaining the specification that minimizes the Bayesian Information Criterion (BIC). (2) Define the lower bound of the bunching window as the leftmost set of two adjacent bins below the threshold where the actual count in each bin exceeds the 95 percent confidence interval of the predicted bin counts from equation (30), 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).

Figure A1: 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 $e_0 = 0.3$.

In Figure 14b, 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 5. A higher elasticity e increases the density at the kink point by raising the total amount of bunching mass (Figure 5a). 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 5b). Thus, the parameter combinations to the southeast of the baseline in Figure 14b correspond to densities with substantially higher density around the kink point, like the tallest density displayed in Figure 14b. 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 14b 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 14b.[32]

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.

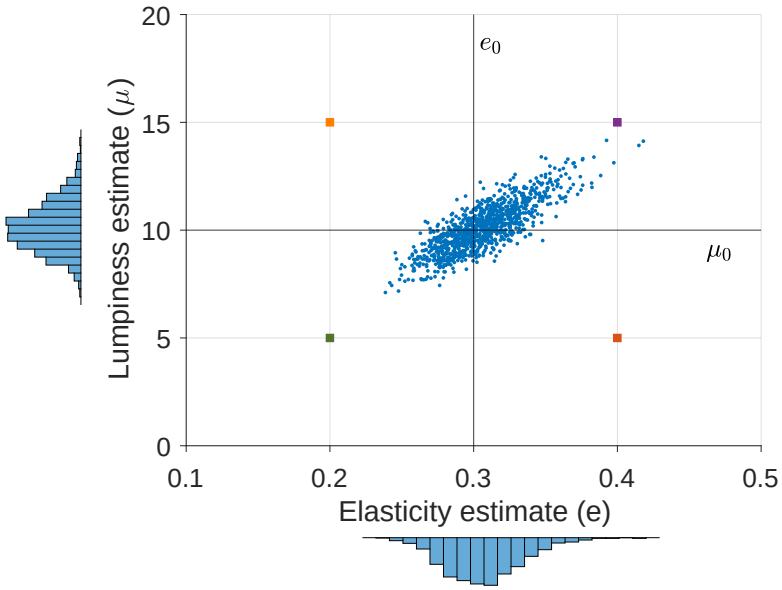
A.4 Alternative specifications for the conventional approach

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

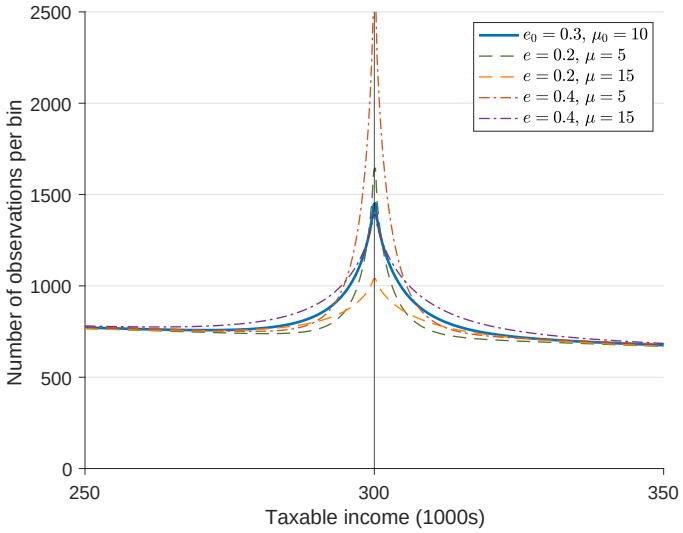
[32] Put differently, the estimator that we propose would find it easier to distinguish between data generated from "low e, high μ" and "low μ, high e" combinations than between "low μ, low e" and "high μ, high e" combinations.

Figure A2: 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 14a. 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).

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 A3a. 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 A3a.

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 A3b. 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 iden-

tical 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 A3a. 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 A3a. 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.

A.5 Comparison to the conventional notch-based bunching estimator

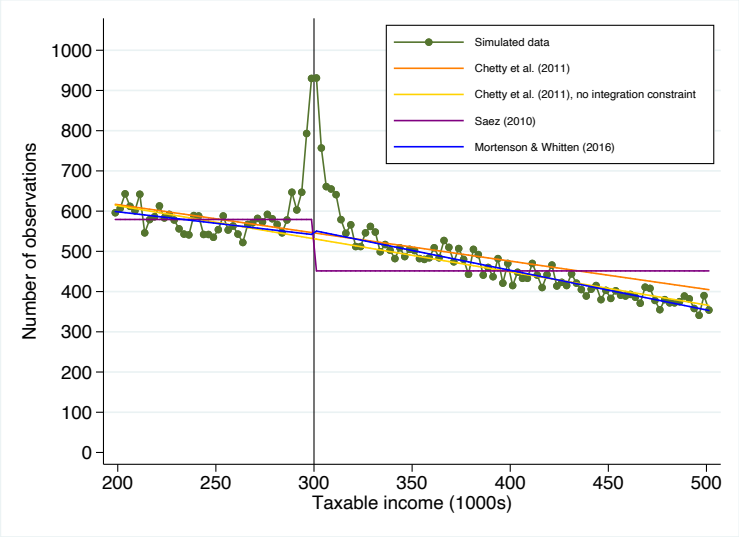
Kleven and Waseem (2013) (hereafter KW) proposes a method for estimating the elasticity of taxable income based on bunching around a notch in the tax schedule. Here we apply that method to simulated data with sparsity-based frictions. Figure A4 plots an income histogram for one simulation round with parameters identical to those in Figure 14, 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.

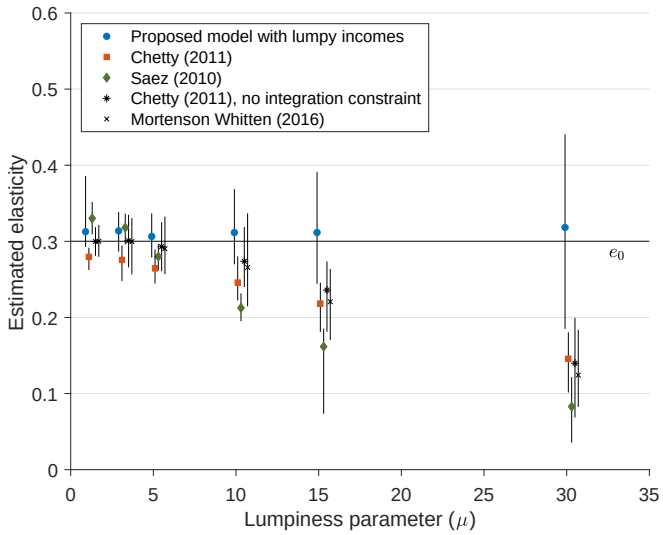
Panel (b) displays the results of applying the KW notch-based estimator. In this model, the presence of taxpayers at incomes that are strictly dominated (i.e., at which post-tax income is lower *and* labor effort—as evidenced by pre-tax income—is higher than at k) is explained by assuming that a share 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

Figure A3: 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.

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.[33] 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).[34]

The resulting elasticity estimate is $\hat{e} = 0.61$—a value that is higher than the elasticity of the data-generating process, $e_0 = 0.3$.[35] 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 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 Figure 5, 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

[33] 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.

[34] 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.

[35] 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.

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.

A.6 Details of South African small business corporations

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 A5 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:[36]

- The business is a close corporation, co-operative, private company or personal liability company.[37]
- 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.[38]
- Investment income and income from rendering personal services may account for a maximum of 20 percent of all receipts, accruals and capital gains.[39]

[36] 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`.

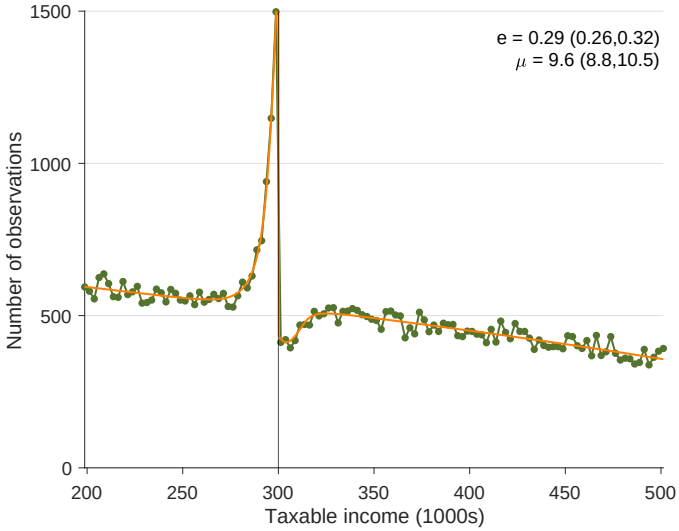
[37] A close corporation is a firm that was required to have 10 or less owners. After 2019, new companies could no longer incorporate as close corporations, but previously registered close corporations could maintain this form.

[38] 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.

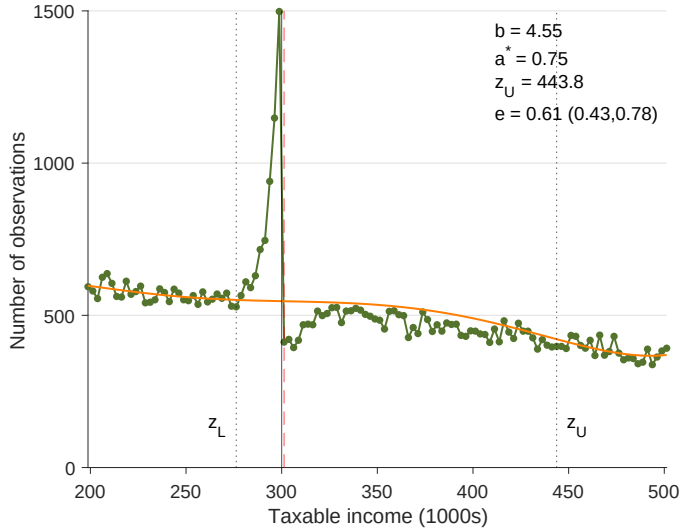
[39] Personal services refer to any company services performed personally by any person who holds an interest in that company when that company employs less than 3 employees. In this scenario, the tax authority deems the income being generated to be a function of the personal skill of that individual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

Figure A4: Elasticity estimation based on Kleven and Waseem (2013)

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

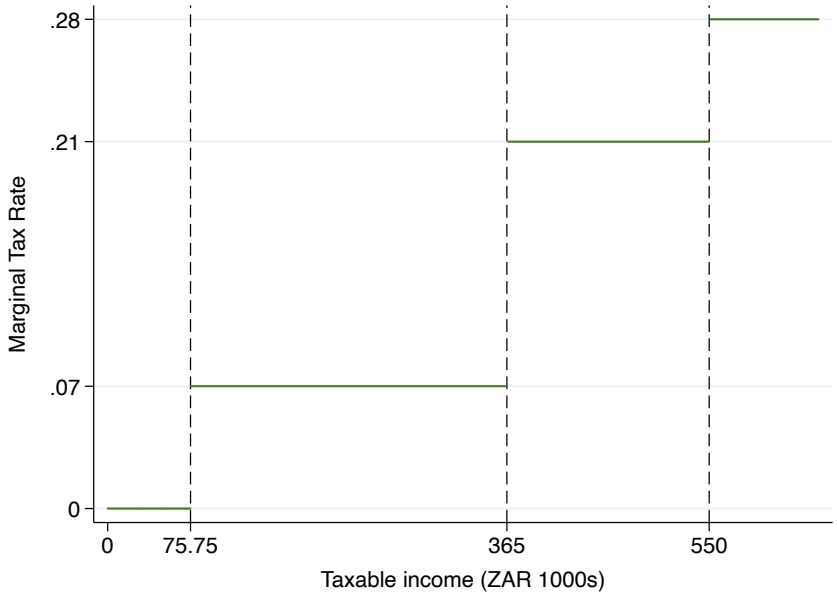


(b) Estimation based on Kleven and Waseem (2013)



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 14. 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 A5: Tax Schedule for Small Business Corporations in 2018



This figure shows the marginal tax rate schedule for small business corporations in South Africa in 2018. The horizontal axis is measured in South African rand (ZAR), and vertical dashed lines indicate bracket thresholds where marginal tax rates change.

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

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.

companies.[40]

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

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

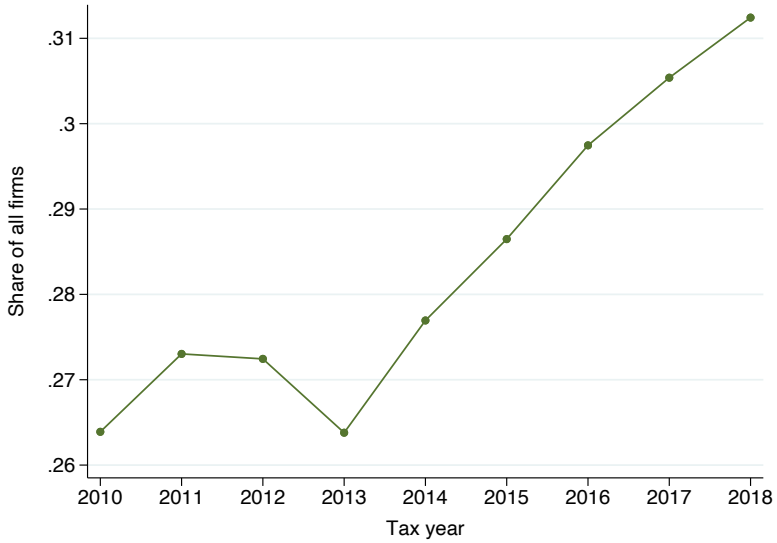
Company Type	Non-SBC	Non-SBC Size Matched	SBC
Turnover (in R'000)	121,249.1	3,028.22	2,628.6
	(521,791.0)	(4,275.89)	(3,531.4)
Expenses (in R'000)	33,713.74	1,684.13	1,244.78
	(281,924.3)	(2,632.84)	(1,785.2)
Assets (in R'000)	73,928.56	3,422.88	1,264.86
	(658,585.4)	(16,459.2)	(2,344.89)
Liabilities (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.

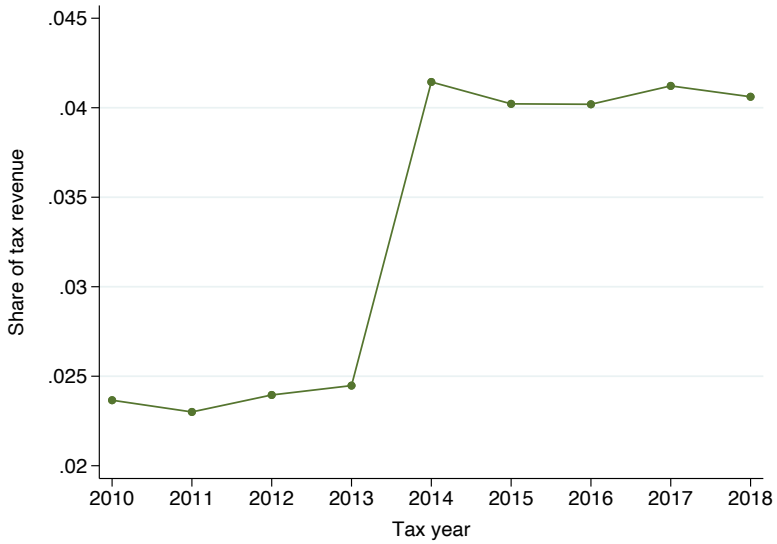
The share of SBCs has risen over time. Figure A6a shows a clear upward trend since 2013, with SBCs accounting for over 31 percent of all tax filings in 2018. While large in number, the contribution of SBC tax revenue to total corporate income tax revenue, as shown in Figure A6b, is more modest, with 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.

Figure A6: 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.